university of groningen

faculty of science and engineering

Approximation of homogeneous networks using Exponential Random Graph Models

Master Project Mathematics

July 2018

Student: G. Ceoldo

First supervisor: Prof.dr. E.C. Wit

Second supervisor: Prof.dr. W.P. Krijnen

I would like to express a special thanks to professor Ernst Wit for his guidance in this research project, to my second supervisor prof. Wim Krijnen for the suggestions on how to improve my thesis, and to Spyros Balafas for having motivated me multiple times.

Abstract

Exponential random graph models are the main parametric approach to do statistical inference in network data. This thesis is an introduction to the rationale behind exponential random graph models, which is based on maximization of Shannon entropy as inference procedure, properties of exponential families and theory of random graphs. The research question is to find an exponential random graph model that can approximate every network distribution invariant under relabelling of vertices. In this model, the first moments of the eigenvalue distribution of the network are used as sufficient statistics. The moments of the eigenvalue distribution are equal to the total number of closed walks of different lengths, multiplied by the size of the network.

The approximation works because of the geometrical properties of the log-density space of exponential families. These properties allow a decomposition of the space in term of information explained by the approximating model, and residual information orthogonal to the previous one. The main result is the explicit computation of the relative entropy of the real network distribution with respect to the approximating ERGM. If the size of the homogeneous network under study is fixed, the ERGM with number of close walks as sufficient statistics can approximate arbitrarily well the real distribution, if enough moments are used.

Contents

1 Introduction **3**

2 Maximum Entropy and Exponential Families **9**
2.1 Measure of Information and Uncertainty 9
2.2 Maximum Entropy Principle 11
2.3 Exponential Family of Distribution 14
2.4 Geometry of Exponential Families 19
2.5 Minimum Information Projection 22

3 Random Networks **27**
3.1 Adjacency Relations 28
3.2 Descriptive Statistics 29
3.3 Models for Random Networks 34

4 Exponential Random Graph models **38**
4.1 Joint and Conditional Distributions 39
4.2 Sufficient statistics and Local Dependencies 42
4.3 Markov Random Graphs 46
4.4 Theoretical and Computational Problems 48

5 Exponential Random Graph Models with Spectral Statistics **54**
5.1 Higher Order Local Dependencies 54
5.2 ERGM as Parametrized Centrality Measure 60
5.3 Final Remarks 61

Bibliography **65**

List of symbols

$H(P)$	Entropy of P
$D(P\|\|Q)$	Relative entropy of P from Q
P_θ	(Exponential) family of distributions, parametrized by θ.
$\mathcal{P}_\mu$	Set of probability distributions, with moments μ.
$\mathbb{P}(A)$	Probability of the event A
$\mathbb{E}(X)$	Expected value of the random variable X
$\mathbb{V}(X)$	Variance matrix of X
$V(G)$	Vertices of the graph G
$E(G)$	Edges of G
$D(G)$	Dependence graph of G
$N(i)$ $[N(ij)]$	Neighbours of the vertex i [edge ij]
C_j	j-dimensional cycle
S_j	j-star
K_j	j-dimensional complete graph
0_j	j-dimensional empty graph
$\widetilde{\mathscr{C}}(G)$	Set of maximal cliques of G
$\mathscr{C}(G)$	Set of different maximal cliques of G
$\mathscr{S}_c(G)$	Subgraph of G correspondent to the clique $c \in \mathscr{C}(D(G))$
$W_j(G)$	Number of closed walk with j steps
$\widetilde{W}_j(G)$	$W_j(G)/j!$
$\mathcal{G}_n$	Space of simple graphs with n vertices
$\mathbf{A}(G)$	Adjacency matrix of G
d	Degree sequence
$\tilde{d}(x)$	Degree distribution
λ	Eigenvalue sequence
$\tilde{\lambda}(x)$	Eigenvalue distribution

Chapter 1

Introduction

Many complex objects can be represented by a fixed set of points joined by ties. The object under analysis is therefore a graph, in this thesis the interest is when the set of vertices is fixed and the ties are random accordingly to some statistical models. This is a probability distribution defined in the space of graphs and, if specified correctly, the distribution can be interpreted as information on the system under study.

Uncertainty in the Specification of a Statistical Model

Chapter 2 starts with formal definitions of information and uncertainty of a probability distribution. In short, an event is a particular result of a random experiment, the information of the event is defined as the amount of knowledge we gain, after the event is observed. An event has higher information than another if it's less likely and the knowledge gained by two independent event is the sum of the individual information. The distribution that model the experiment specifies the probability of each event, so the distribution can be defined coherently with respect to the information gained in every possible outcome of the experiment.

The expected information of a (distribution of a) random variable is called entropy, it is a measure in the space of distributions that can describe the experiment. The entropy can be used to choose a particular distribution as model of the data. Specifically, distributions with higher entropy are more uncertain, as the events contain more information, on average. Thus, if we choose a distribution with higher entropy than another one, we are imposing less assumptions on the statistical model, as the data (outcomes of the experiment) lead to an higher average knowledge gain. More precisely, as we define information in term of knowledge gain, we should fix a prior distribution that represents the knowledge that we have on the object under study before seeing the outcome. In this set a good measure of information is the relative entropy of the specified distribution with respect to the prior. The entropy corresponds to the inverse of the relative entropy when the prior is noninformative, so when it represent

the situation in which before seeing an outcome we don't know anything on the system under study.

The main approach is to choose as statistical model the distribution that maximizes the entropy with constraints that fix the expected values of the sufficient statistics of the distribution. The same approach can be derived also axiomatically, the maximization of the entropy (or more generally, minimization of the relative entropy) is the only approach that assures a consistent specification of the statistical model, in term of some axioms that a good inference procedure should satisfy. The distribution obtained by maximization of entropy belongs to an exponential family, generated by the sufficient statistics whose expected value has been fixed in the constraints of the maximization algorithm. The exponential family is one of the main approaches to define statistical models. In our case it is particularly useful because in this family the sufficient statistics contain all information of the distribution. Therefore, the exponential family can be seen as a distribution in the space of sufficient statistics or configurations, different distributions with same sufficient statistics are indistinguishable, therefore they carry the same information.

The exponential families are most often used without consider their interpretation in term of information, as they have optimal mathematical properties in many cases, because of the tractability of their log-density. For example, the log-likelihood ratio between two distributions in the same family is a linear function of the sufficient statistics, and it can be decomposed in the sum of the log ratios between the two initial distributions and a common distribution in the family. This decomposition can be extended in term of relative entropy. More interestingly to our problem, the sufficient statistics lie in a vector space, so there are explicit solutions for approximating a family with another one generated by sufficient statistics in a linear subspace of the original one. For continuous distributions the vector space is infinite dimensional, however many distributions can be written in exponential form using an infinite dimensional basis of continuous functions (for example orthogonal polynomials) that spans the configuration space.

The exponential family can be used also as approximation of distributions that does not belong to the family. Consider the set of distributions defined fixing the expected values of some functions of the data. The information projection of a exponential family of distributions to the set defined before is the only distribution that belong both to the set and the exponential family, therefore it's the natural approximation of a distribution that does not belong to the family. The estimate of the distribution is in the same family as the approximation. The information projection allows the decomposition of the information loss, which is the sum of the approximation error and estimation error. The first is the information loss of approximating a distribution with the closest one that belong to the exponential family, the second is the difference between the information projection and the estimate using the real data.

Networks as Representation of Complex Systems

If the system we are interested is a collection of objects joined by connections between them, the object under study can be represented using a network. The connections are random, therefore a probability distribution is defined in the space of possible networks, or equivalently in the space of possible adjacency matrices. Pattern of ties in the observed networks are called subgraphs, many of them are associated with descriptive statistics of the adjacency matrix (i.e. information on the observed network).

The most used descriptive statistics is the degree distribution, which specifies the proportion of vertices with a given degree. The first moment of this distribution is the average degree. The relation between this statistics and the size of the network is very important. In fact, for most real large networks, the average degree grows proportionally to the size of the network. A ranking of the vertices, based on their importance in the network structure, is called centrality The simplest one is the degree centrality, in which the importance of a vertex is measured in how many connections it has.

It is possible that the degree distribution does not represent all information contained in the network. However, for homogeneous networks, i.e. graphs in which the vertices are unlabelled, all information of the network is in its eigenvalues, because the information on the "location" of the vertices in the graph is in its eigenvectors. If the size of the network is fixed, the eigenvalue distribution is discrete and it has a bounded support. Therefore, it can be identified by its moments, which are associated with the expected number closed walks in the network. The total number of closed walks with a given length is a linear combination of the counts of some subgraphs. "Very complicated" subgraphs affect only the counts of very long closed walks, therefore the first moments of the eigenvalue distributions are function of small subgraphs, which usually are considered to be more important in the network structure.

Similarly to the degrees, a one dimensional sufficient statistics can be defined as a weighted sum of number closed walks with different length. The most important one is the Estrada index and it's related with the statistical mechanical properties of the network. We can also define a centrality measure that quantifies the importance of a vertex based on how many closed walks pass through it. The most important one is called Estrada centrality and depends on the eigenvalue distribution, as the degree centrality depends on the degree distribution.

Descriptive statistics can be used to define a model for random graphs. There are two main approaches to do so. The first one specifies an algorithm for generating a random graph with given property, for example the average degree, or a particular degree distribution. The simplest example is the Erdös-Rényi random graph, in which the only parameter is the constant probability of a tie between every possible pair of vertices. Almost all models defined in this way are extensions of the Erdös-Rényi graph. In principle, if the model is correctly specified, should be possible to compute the expected values and distributions of some descriptive statistics defined before. However, it's usually

nontrivial how to do so, the other main approach starts from the choice of which sufficient statistics are important, and then the distribution is specified as the one maximize our uncertainty, given the expected value of the important sufficient statistics.

Network Distributions with Maximum Entropy

The second main approach for defining models for random networks is related to thermodynamics and statistical inference. The expected values of the sufficient statistics that are considered to be important in the model are fixed. Then, the entropy is maximized leading to a random graph with distribution that belong to the exponential family. The generative algorithm in this case is a (thermodynamic) stochastic process governed by parameters, and the network is an observation of this process in a fixed time. In Huang (1963) the parameters of the family are defined to be "measurable macroscopic quantities associated with the system". We need to assume that when the network is observed, the process is in a sort of thermodynamic equilibrium, therefore the distribution indexed by the macro parameters is close to the maximum entropy distribution.

Maximization of entropy leads to a distribution that belong to the exponential family generated by the sufficient statistics whose expected value is fixed in a one-to-one relations with the macro parameters. If the system under study is a network, this distribution is called Exponential Random Graph Model. These models are useful in statistical inference because it is possible test the significance of the parameters, as the distribution of their associated sufficient statistics is known. The inference is done by Markov Chain Monte Carlo, because it mimics the generative process in which ties are created and eliminated with time. This method is necessary because in this way the estimates are averages of multiple networks that came from a process in thermodynamic equilibrium (if the chain is stable).

The Hammersley-Clifford theorem (Besag (1974)) implies that every random network can be written with a distribution that belong to the exponential family. The theorem also specifies which sufficient statistics are associated with non-zero macro parameters and so they are significant in the generative algorithm. These sufficient statistics are particular subgraphs, and when the network is homogeneous their "location" in the graph is not important, so the parameters can be associated with the counts of these subgraphs in the network. The most used exponential random graph model are variants of the Markov random graph. For this model the sufficient statistics are the degree distribution and the number of triangles. Thus the Markov random graph is the distribution with higher entropy when the expected number of triangles and the expected degrees of the vertices are fixed. The downsizes of these approach are caused by the instability of the sufficient statistics of the model in canonical form. Moreover, this model does not specifies how the network grows, therefore consistency of the estimates can be an issue.

Approximation of Homogeneous Network Models

Our original research question was to find an Hamiltonian such that we can approximate every network distribution using an exponential random graph model. The starting point is the Hammersley-Clifford theorem, because every network distribution can be written as an ERGM, the theorem specifies also the form of the Hamiltonian, which is a linear combination of subgraph counts. The Hamiltonian can be simplified considerably if we consider only homogeneous networks. In this thesis an homogeneous network is defined to be a random graph with distribution invariant under relabelling of the nodes (sometimes homogeneity is defined in a different way, see Estrada (2012) chapter 9 for example). Markov random graphs do not use all information contained in the network, because the degree distribution does not contains all information.

The homogeneity condition implies that the eigenvalues contains all information of the network. Therefore, our idea has been to use as Hamiltonian a linear combination of the first moments of the eigenvalue distribution, this setting offers an interpretation of the parameters in term of number of closed walks. More interestingly, the number of closed walks are linear combinations of the subgraph counts specified by the Hammersley-Clifford theorem. Therefore the model introduced in the last chapter is an approximation of every possible homogeneous network model, and the approximation is linear in the space of sufficient statistics.

Using the affine geometry of exponential families, the approximation can be evaluated completely in the log-space. Therefore, the quality of the approximation depends on how well a linear combination of number of closed walks can approximate a linear combination of subgraph counts. The sufficient statistics of the Hamiltonians are discrete functions of the networks, in particular they are basis of spaces of log-densities, which are finite dimensional. The log-density space of the approximated distribution is than a lower dimensional subspace of the original one. The affine geometry implies that vectors in the original space can be decomposed as a sum of a vector in the approximating space and an orthogonal residual. Therefore, the information loss is orthogonal to the space of closed walks considered in the Hamiltonian. However if enough sufficient statistics are used, the real and approximating model are equivalent.

For non homogeneous networks, like the ones with community structure for example, these results are not directly applicable. In fact, even with Markov dependency, modelling the community structure with an exponential random graph models demands many parameters. Therefore, the original research question has been solved partially. However, different Hamiltonians can be combined as sufficient statistics of the network. Thus, the community structure (or other kind of inhomogeneities in the network) can be modelled with an Hamiltonian derived by Markov dependency for example, and the residual information can be modelled using the closed walks Hamiltonian. With this example, the only information which is not explained is the non-homogeneous one orthogonal to the information fitted with Markov dependencies.

Notes on the Literature

The main inspiration for this work has been Estrada (2012), which is a great introduction to random networks, often from the statistical viewpoint of how to "extract" information from network data, mostly in term of descriptive statistics. Complex graphs contain a lot of information, and the book is particularly useful because a network can be analysed from many different perspectives. The maximization of entropy as inference procedure starts with Shannon (2001) (originally published in 1948), however my thesis is based mainly on Jaynes (1982), his work connected the Shannon's interpretation of the entropy in term of information with the physical one developed previously by Boltzmann and Gibbs. Distribution with maximum entropy are exponential families, the idea of exploit the affine geometry of the log-space of these distributions is inspired by Jørgensen and Labouriau (2012), chapter 1. Lastly, the most important references for the discussion on exponential random graph models are Newman (2010), section 15.2, Robins et al. (2007), Snijders et al. (2006) and Frank and Strauss (1986). I tried to emphasize the connection between exponential random graph models and the theory developed in chapter 2, because I think that both the rationale and the mathematics behind these models is based mainly on their thermodynamic interpretation.

Chapter 2

Maximum Entropy and Exponential Families

Introduction

Parametric statistical inference starts with the definition of the model that generates the data. When the distribution of the random experiment is unknown, its specification involves some assumptions in the generator mechanisms. Other than the constraints on the distribution, given by the set up of the random experiment, we would like not to impose other assumptions in the distribution, but still use a parametric statistical model. If the experiment allows us to gain some information on the system under study, we are imposing less assumption if we choose as model the distribution with maximum uncertainty, because the data gives, on average, more information. The starting point is the definition of a measure of information which, if maximized, leads to a distribution in the exponential family.

2.1 Measure of Information and Uncertainty

A result of a random experiment can be associated with the information gained to the system under study, before and after the experiment is performed. Formally, *information* is defined as the knowledge gained by the result of a statistical experiment, identified by a probability distribution defined in the object under study. The information is on the *macro-properties* of the system, governed by real *parameters*. When we have knowledge about important combinatorial properties of the system under study, it is non-trivial assign probability to various occurrences in a consistent way.

Usually in this thesis will be considered experiments modelled by a finite dimensional *discrete probability distribution* X with possible values $x_1, ..., x_N$ each of them with probability $p_1, ..., p_N$ such that $p_i > 0$ and $\sum_{i=1}^{N} p_i = 1$. We

can associate the information of this experiment by the *uncertainty* we have in defining the probability distribution. In this way the concept of information is associated to the random variable to model the experiment, in particular it's equivalent of the uncertainty we have in choosing a specific random variable. The following axioms allows to derive mathematically a measure of information (or uncertainty) of the discrete random variable $p_1, ..., p_N$:

- If exist i such that x_i has probability one (and consequently all other possibilities have probability 0), then the experiment has zero uncertainty (there is no information gain on performing the experiment).
- If $p_i < p_j$, than the event $X = x_i$ is less uncertain than $X = x_j$ (the first event carries more information).
- If the events $X = x_i$ and $X = x_j$ occur independently, then the uncertainty (information) of the joint occurrence, is the sum of the uncertainties (information) of the singular events.

The three axioms force the mathematical form of the measure of information or uncertainty f. In fact if A and B are independent events, with associated information $I(A)$ and $I(B)$, then

$$I(A \cap B) = I(A)I(B) \quad \Longleftrightarrow \quad f(\mathbb{P}(A \cap B)) = f(\mathbb{P}(A))f(\mathbb{P}(B)), \tag{2.1}$$

which implies

$$I(A) = -b \log \mathbb{P}(A), \tag{2.2}$$

where b is a constant greater than 0 that fix the unit of measurement. If $b = 1/\log 2$ the information is measured in *bits*, if $b = 1$ in *nats*. In this thesis is always assumed $b = 1$.

The random variable is a model for the experiment, its uncertainty is the *expected information gain* or *(Shannon) entropy*, defined as the average information of the events modelled by the random variable:

$$H(P) = \int I(\omega) dP(\omega) = \mathbb{E}_P(I(X)), \tag{2.3}$$

in the case X has discrete distribution P the entropy is

$$H(P) = -b \sum_{i=1}^{N} p_i \log p_i = -b \sum_{i=1}^{N} \mathbb{P}(X = x_i) \log \mathbb{P}(X = x_i). \tag{2.4}$$

If we have *prior information* on the experiment, it's possible to define a measure of uncertainty with respect to the random variable that represents prior knowledge of the system, which will be denoted as Q. This is called *relative entropy* or *Kullback-Leibler divergence* from Q to P is

$$D(P||Q) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} = \sum_{i=1}^{N} \mathbb{P}_P(X = x_i) \log \frac{\mathbb{P}_P(X = x_i)}{\mathbb{P}_Q(X = x_i)}. \tag{2.5}$$

D is not a distance between P and Q because is not symmetric, however it is equal to 0 if and only if P and Q are equal with probability 1. If Q is the uniform distribution, then $D(P||Q) \propto -H(P)$ which means that the relative entropy of P with respect to the uniform is the inverse of the Shannon entropy of P.

The relative entropy can be interpreted also in a frequentist way. In this framework, P is the real distribution of a random experiment, Q is the model specified by the statistician. $D(P||Q)$ is the expected value of the difference in term of information between the real distribution and the model, the expected value is computed with respect to P. Therefore $D(P||Q)$ is a measure of information loss by approximating P with Q. The approximation Q can be seen as an estimation of P in information sense. Usually when Q is an estimator of P is denoted by $\widehat{P}$. The use of entropy as information measure is introduced in Shannon (2001) (originally published in 1948), my presentation is based mostly on Martin and England (2011), chapter 2.

2.2 Maximum Entropy Principle

In the previous section, the Shannon entropy is introduced as measure of uncertainty. Each distribution corresponds to the average information gained by the experiment, we may want to be as uncertain as possible in the definition of the model, so that the data lead to higher knowledge, on average. Jaynes (1982) consider the maximization of entropy as an inference procedure optimal if the data are measured without error. In this case all uncertainty that we have in the specification of the model, depends only on the fact that the real distribution is unknown. The Shannon entropy is a measure in the space of probability distributions that favours models which can be realized in more ways, by the combinatorial properties of the sample space. In some cases there are concentration inequalities that run out models with distributions far from the one with maximum entropy. Conversely, if the data have noise, maximization of entropy does not work well. In this case it's useful to exploit properties and symmetries of the noise.

If we have no information on the choice of the probability distribution, there are no restrictions on p_i. The maximization of entropy choose P that maximizes

$$\max_P H(P) = \max_{p_1,\dots,p_N} -\sum_{i=1}^{N} p_i \log p_i. \tag{2.6}$$

If there are no restrictions, the solution of the maximization problem is the uniform distribution $p_i = 1/N$, which is the one that carries most uncertainty. Equation 2.6 is called *Principle of indifference*. It can be used to choose an appropriate prior distribution as, if the random variable is bounded (like in our example) the solution is the uniform measure. The prior chosen in this way is called *noninformative prior*. However, when information is available, different distributions emerge as solution of the problem 2.6.

Maximum Entropy with Linear Constraints

The problem is to choose a "good" probability distribution to model the random experiment. This choice has to be coherent with the information we have on the experiment. If there is none, maximization of entropy gives as choice the most uncertain (less informative) distribution: the uniform. Consider the case in which information is available, in form of expectations of functions $f_j(X)$. The problem 2.6 becomes

$$P^* = \underset{P}{\operatorname{argmax}}\ H(P) \quad \text{s.t.} \quad \sum_{i=1}^{N} p_i f_j(x_i) = \mu_j, \quad \sum_{i=1}^{N} p_i = 1, \tag{2.7}$$

where $X \sim P$, $p_i = \mathbb{P}(X = x_i)$, $f = (f_1, ..., f_m)$ and $\mu = (\mu_1, ..., \mu_m)$ is the information available in term of the moments of f_j.

The maximization is done with $m+1$ Lagrange multipliers, α_0 for the normalization constraint, and $\alpha_1, ..., \alpha_m$ for the constraints on the expected values. The *Lagrangian* is

$$\mathcal{L}(P, \alpha) = -\sum_{i=1}^{N} p_i \log p_i - \alpha_0 \left(\sum_{i=1}^{N} p_i - 1 \right) - \sum_{j=1}^{m} \alpha_j \left(\sum_{i=1}^{N} p_i f_j(x_i) - \mu_j \right), \tag{2.8}$$

the derivatives with respect to p_i and α_j are

$$\begin{aligned} \frac{\partial}{\partial p_i} \mathcal{L}(x, \alpha) &= \log p_i - \alpha_0 - p_i \sum_j \alpha_j f_j(x_i), \quad i = 1, ..., N \\ \frac{\partial}{\partial \alpha_j} \mathcal{L}(x, \alpha) &= \sum_{i=1}^{N} p_i f_j(x_i) - \mu_j, \quad j = 1, ..., m \\ \frac{\partial}{\partial \alpha_0} \mathcal{L}(x, \alpha) &= \sum_{i=1}^{N} p_i - 1. \end{aligned} \tag{2.9}$$

Equating the derivatives to 0, gives the solution P^* of the optimization problem which is the distribution

$$p_i^* = \mathbb{P}(X = x_i) = \frac{1}{Z(\alpha)} e^{\sum_{j=1}^{m} \alpha_j \cdot f_j(x_i)} = e^{\alpha \cdot f(x_i) - F(\alpha)}, \tag{2.10}$$

where $Z(\alpha)$ and $F(\alpha)$ are called respectively *partition function* and *free energy* (or *log-partition function*), defined as

$$Z(\alpha) = \sum_{i=1}^{N} e^{\alpha \cdot f(x_i)}, \quad F(\alpha) = \log Z(\alpha). \tag{2.11}$$

The parameters $\alpha_1, ..., \alpha_m$ gives the explicit relation between the partition func-

tion and the information available, specifically

$$\mu_j = \mathbb{E}_{P^*}(f_j(X)) = \frac{\partial}{\partial \alpha_j} F(\alpha_1, ..., \alpha_m). \tag{2.12}$$

Note that α_0, the Lagrange multiplier associated with the normalization constant is included in the partition function, so is not a parameter of the distribution P^*.

The maximum entropy is

$$H(P^*) = \alpha \cdot \mu - F(\alpha), \tag{2.13}$$

if x is an n-dimensional i.i.d. sample from X and the "theoretical" moment μ_j is replaced by its empirical version $\hat{\mu}_j = \sum_{i=1}^{n} f_j(x_i)$, then the maximal entropy with sufficient statistics $\hat{\mu}_j$ is the *log-likelihood*, if taken as a function of α. This connect the approaches of maximum entropy and maximum likelihood.

Axiomatic Derivation

The former argument motivates the choice of the distribution of maximum entropy as it is optimal in term of information and uncertainty. However, the approach of Shore and Johnson (1980) motivates 2.7 because it's the only inference rule that guarantee a self-consistent choice P^* (Pressé et al. (2013)). Define the function $\widetilde{H}(P, Q)$ which depends on the distribution P and the prior Q. The chosen distribution P^* is the solution of the optimization problem

$$\max_P \widetilde{H}(P, Q) \quad \text{s.t.} \quad \mathbb{E}_P(f_j(X)) = \mu_j, \text{ for } j = 1, ..., m. \tag{2.14}$$

Their approach lead to the optimal function $\widetilde{H}(P, Q)$ such that if maximized with respect to the information constraints, lead to "optimal" inference in some sense.

A *self-consistent* inference procedure respects the following axioms, the maximum of 2.14 should be:

- unique,
- invariant with respect to change of coordinates,
- subset-independent (the inference leads to the same result if carried on in the whole set of data or separately in independent subsets of the system, and then combined),
- system-independent (the inference from data that comes from independent systems, can be carried on separately and then combined).

The axioms specify the form of $\widetilde{H}(P, Q)$, which has to be equal to

$$\widetilde{H}(P, Q) = -b \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} = -bD(P||Q), \tag{2.15}$$

(Shore and Johnson (1980), Pressé et al. (2013)) or, if Q is uniform,

$$\widetilde{H}(P) = -b \sum_{i=1}^{N} p_i \log p_i. \tag{2.16}$$

So the optimal inferred distribution is the one that maximizes the entropy, subject to the constraints, or, if prior information Q is available, minimizes the relative entropy from Q to P.

Note that in this derivation the fact that $H(P)$ and $D(P||Q)$ are measures of information and uncertainty is not used at all. Despite so, if we are interested in this interpretation, the axioms are reasonable. For example, the axioms of system and subset independence, implies that the inference in the various subsets (or systems) can be combined so that the joint information of the experiment is the sum of the information in the various subsets (or systems). The maximum entropy principle is viewed as the only inference rule that respects the axioms. Self-consistency is a different concept than consistency in statistics, is related to assign the distribution of the model in a coherent way, while consistency is related to the quality of an estimate of the "unknown" parameters or distribution.

These axioms specify also which information constraints are appropriate in the maximization. First, the constraints have to be linear in p_i, they can be either equalities or inequalities. Many functions f_j can be used to define the constraints, for example polynomials are appropriate, which implies that all moments of a distribution function can be used (Pressé et al. (2013)). The various interpretation of entropy are summarized in Giffin (2009), the definition of Shannon, the one I used, is in term of information and generalizes the definitions used before, as is shown in Jaynes (1957).

2.3 Exponential Family of Distribution

In statistics, the choice of a distribution that describes a random experiment is based on assumptions on the properties of the data generated from the experiment. In this section, the exponential family of distribution is derived in the classic statistical approach, and many quantities which was previously defined only for discrete distributions will be extended to other measures.

The space of possible outcomes of the experiment is *sample space* $\mathcal{X}$. This space is enriched with a collection of subsets of $\mathcal{X}$, called *σ-algebra*, denoted by $\mathcal{A}$ and a measure ν over it. The two cases that will be considered are when the sample space is *discrete* or *absolutely continuous*. In the first case, the σ-algebra is $2^{\mathcal{X}}$, the *power set* of $\mathcal{X}$ which consists of all the possible subsets of elements of the sample space, the measure ν is the *counting measure*, which counts the number of elements in $a \in \mathcal{A} = 2^{\mathcal{X}}$. In the second case, is considered the *Borel σ-algebra*, formed by countable unions, intersections and complements of the open sets in $\mathcal{X}$, here ν is the *Lebesgue measure*, the standard measure of volume in an Euclidean space.

With this construction, the sample space $(\mathcal{X}, \mathcal{A}, \nu)$ becomes a *measure space*. The first step in the definition of the probability distribution involves the transformation of the measure ν through an $\mathcal{A}$-measurable function $h : \mathcal{X} \to \mathbb{R}$ which is called *reference function*. h induces to $(\mathcal{X}, \mathcal{A})$ the (not necessarily finite) *base measure* (or *reference measure*) ν_h, which often admits density

$$d\nu_h(x) = h(x)d\nu(x), \tag{2.17}$$

for $x \in \mathcal{X}$. The measure space $(\mathcal{X}, \mathcal{A}, \nu_h)$ induces a *reference distribution* on $\mathcal{X}$ with density

$$dP_{0,\nu_h}(x) \propto d\nu_h(x) = h(x)d\nu(x). \tag{2.18}$$

P_{0,ν_h} is defined in the *support of the reference measure* ν_h is the closed set

$$\{x \in \mathcal{X} : \nu_h(x) > 0\} \quad \text{or} \quad \{x \in \mathcal{X} : \exists \epsilon > 0 \text{ s.t. } \nu_h(N_\epsilon(x)) > 0\}, \tag{2.19}$$

for the discrete or continuous case respectively, $N_\epsilon(x)$ is the neighbourhood of x with radius ϵ. It can be assumed $h > 0$ so that only ν specifies the support. Denote with K_ν the closure of the convex hull of the support.

The reference distribution P_{0,ν_h} can be expanded to a *family of probability distributions*. Let α a m-dimensional vector of parameters, $f = (f_1, ..., f_m) : \mathcal{X} \to \mathbb{R}^m$ an $\mathcal{A}$-measurable function. The vector $\alpha \in A$ parametrizes the *exponential family of distributions* with base measure ν_h and *sufficient statistics* $f_1, ..., f_m$. A fixed value α^* in the *parametric space* A, induces the probability distribution with density

$$dP_{\alpha^*;f,\nu_h}(x) \propto \exp(\alpha^* \cdot f(x))d\nu_h(x) = \exp\left(\sum_{i=1}^{m} \alpha_i^* f_i(x)\right) h(x)d\nu(x). \tag{2.20}$$

The exponent term $\alpha^* \cdot f(x)$ sometimes is called *Hamiltonian*. Apart from the base measure, which is common to all densities in the family, every specific distribution depends on the data only through the sufficient statistic f, which is a low dimensional summary of the data. In particular, the density is determined by linear combinations of $f(x)$ the main characteristics of the data x.

To obtain a probability measure, $dP_{\alpha^*,;f,\nu_h}$ has to integrate to 1 over its domain. We can standardize the measure computing

$$F(\alpha) = \log \int \exp(\alpha \cdot f(x))h(x)d\nu(x) = \log \mathbb{E}_{\nu_h}(\exp(\alpha \cdot f(X)), \tag{2.21}$$

which is the *log-Laplace transform* of the reference measure, called also log-partition function, or free energy. The family is defined when F is finite and the parametric space can be chosen as

$$A = \{\alpha : F(\alpha) < \infty\} \subseteq \mathbb{R}^m, \tag{2.22}$$

which is called *natural* or *canonical parametric space*.

It's possible define a *parametrization function* such that $\beta = \beta(\alpha)$, so that β is the new vector of parameters. If this function is the identity, then α is called *natural* or *canonical parameter*, and similarly, the *canonical exponential family* has densities

$$dP_\alpha(x) = \exp(\alpha \cdot f(x) - F(\alpha))h(x)d\nu(x). \tag{2.23}$$

Equation 2.22 gives the main advantage of using the natural parameter, in this case the parametric space A coincides with the set of values such that $F(\alpha)$ is finite, i.e. P_α is a probability distribution for all α. The model 2.23 is equivalent of the one obtained with the transformation

$$\beta = \beta(\alpha) = \mathbf{B}\alpha + \beta_0 \quad \text{and} \quad y = \mathbf{B}^{-1}f(x) + y_0, \tag{2.24}$$

where $\mathbf{B}$ is nonsingular, (Brown (1986), proposition 1.6). The exponential family is therefore invariant with respect to linear transformations of the sufficient statistics. Starting from 2.23, every choice of $\mathbf{B}$, β_0 and y_0 gives a canonical exponential family.

If K_ν, the closure of convhull(supp(ν)) is bounded, and the sufficient statistics are bounded as well in ν_h, then $F(\alpha) < \infty$ for all α in $\mathbb{R}^m$. However, some values of α define a distribution with expected values of sufficient statistics in the boundaries of K_ν, this is mostly a problem of discrete distributions. Note that some values in the boundaries of K_ν are not in supp(ν), because the last one is not necessarily a convex set. For these α, the distribution represent a model in which the expected value of the sufficient statistics is not in the support of the family, so even though the model still makes sense theoretically, it's useless for describing the system under study.

These problems happens especially when the distribution is discrete. in these cases is useful to reparametrize the model through the non-linear function $\mu : A \to M \subset \mathbb{R}^m$ such that

$$\mu(\alpha)_j = \frac{d}{d\alpha_j}F(\alpha), \tag{2.25}$$

for $j = 1, ..., m$. The new parameters are the expected values of the sufficient statistics, they are called *mean value parameters* and they can be defined when $\mu(\alpha)$ is one-to-one, with inverse function $\alpha = \alpha(\mu)$. This choice gives a more direct interpretation of the parameters, as $\mu = \mathbb{E}(f(X))$. The parametric space $M = \mu(A)$ is usually a more complicated subset of $\mathbb{R}^m$ which is equivalent to

$$M = \text{int}(K_\nu) \quad \text{or} \quad M = \text{rint}(K_\nu), \tag{2.26}$$

if the family is minimal or not respectively. Thus the expected value of the sufficient statistics belongs to the *interior* of the support of the base measure when the family is minimal. If is not minimal, then M is the *relative interior* of the support, which is the interior of an affine subspace of K_ν.

If an n dimensional sample from X is available, $P^* = P_{\alpha^*} \in P_\alpha$ can be estimated with $\widehat{P} = P_{\hat{\alpha}} \in P_\alpha$. The parameter $\hat{\alpha}$ is the *maximum likelihood*

estimator of α, which is the value that solves

$$\mathbb{E}_\alpha(f_j(X)) = \frac{1}{n}\sum_{i=1}^{n} f_j(x_i), \tag{2.27}$$

for $j = 1, ..., m$. Therefore the estimator can be written as

$$\hat{\alpha} = \alpha(\hat{\mu}), \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f_j(x_i), \tag{2.28}$$

where $\alpha : M \to A$ is the parametrization function from the mean value parameter to the canonical parameter. The maximum likelihood estimator exists if and only if $\hat{\mu} \in \text{int}(K_\nu)$ (Jørgensen and Labouriau (2012), theorem 1.18). If the family is not minimal the estimator does not exist as the family is not *identificable*, so there are infinite parameter vectors that solves 2.27.

In the previous section the exponential family has been derived with maximization of entropy. The same approach applies also when the sample space is not discrete and bounded. For general probability distributions P and Q with densities p and q absolutely continuous with respect to the measure ν_h, the entropy is

$$H(P) = -\int p(x)\log p(x) d\nu_h(x), \tag{2.29}$$

with $d\nu_h(x) = h(x)d\nu(x)$. The Kullback-Leibler divergence is

$$D(P||Q) = \int p(x)\log\frac{p(x)}{q(x)} d\nu_h(x). \tag{2.30}$$

When $h(x)$ is constant and ν is the counting measure these quantities correspond with the ones in equation 2.4 and 2.5 respectively. If ν is the Lebesgue measure $H(P)$ is called *differential entropy* and $D(P||Q)$ can be interpreted as the divergence between densities p and q. In a general measure space $(\mathcal{X}, \mathcal{A}, \nu_h)$, the exponential family can be derived in the same way as in section 2.2, by maximization of $H(P)$ in equation 2.29 with linear constraints.

The reference function h is defined a priori and influences both entropy and divergence. When P is bounded, the choice that guarantees maximum entropy is when h is constant, both in the discrete or in the continuous case. Note that when this is the case the distribution P conditional to the value of the sufficient statistics is uniform. This makes sense because often the expected value of the sufficient statistics is the only available information on the experiment, therefore different distributions with same sufficient statistics are indistinguishable, in term of information, so the choice $h(x) \propto 1$ guarantees maximal uncertainty. Another coherent choice in an information context is a reference function that depend on x only through $f(x)$, however (for bounded distributions) the entropy is always lower than when is h constant, and the probability of the data is not affected only by the parameters. This approach is not developed further in the thesis, however in the context of exponential random graph models it has been used in Chandrasekhar and Jackson (2014).

Properties of Exponential Families

The log-partition function $F(\alpha)$ is often considered just a normalization constant. However this function specifies all the information in P_α, because of the explicit link between F and the expected values of $f(x)$. The function is differentiable in all values $\alpha \in \text{int}(A)$ and the derivative and the integral sign can be swapped (Brown (1986), theorem 2.2). Moreover the function is strictly convex in its domain A. Derivatives of F can be used to compute moments of the sufficient statistics.

When the distribution is derived by maximization of entropy, the expected values for the sufficient statistics is fixed. However, the relation between the parameters and the sufficient statistics is also in term of variability and correlation between them. The *cumulant generation function* of P_α is

$$\kappa_\alpha(u) = F(\alpha + u) - F(\alpha). \tag{2.31}$$

Therefore, when α is fixed, the expected value of the sufficient statistics is the gradient of the log partition function:

$$\mathbb{E}_\alpha(f(X)) = \nabla_\alpha F(\alpha) = \left(\frac{d}{d\alpha_j} F(\alpha)\right)_j, \tag{2.32}$$

$j = 1, ..., m$, which is the same as equation 2.12. The variance matrix of $f(X)$ is the Hessian of $F(\alpha)$:

$$\mathbb{V}_\alpha(f(X)) = \mathrm{H}_\alpha F(\alpha) = \left(\frac{d^2}{d\alpha_j d\alpha_h} F(\alpha)\right)_{jh}, \tag{2.33}$$

for j, h in $1, ..., m$.

Two distributions in the same family, differs only in the parameters, in this case the log-ratio and the relative entropy have explicit formulas and decompositions. Fix three possible values of the parameter vector α_0, α_1 and α_2 in A. The log-ratio between P_{α_1} and P_{α_0} is

$$\Lambda_{\alpha_1,\alpha_0}(x) = \log \frac{p_{\alpha_1}(x)}{p_{\alpha_0}(x)} = (\alpha_1 - \alpha_0) \cdot f(x) - (F(\alpha_1) - F(\alpha_0)), \tag{2.34}$$

multiplying both term of the ratio by p_{α_2}, the ratio can be decomposed as

$$\Lambda_{\alpha_1,\alpha_0}(x) = \log \left(\frac{p_{\alpha_1}(x)}{p_{\alpha_2}(x)} \frac{p_{\alpha_2}(x)}{p_{\alpha_0}(x)}\right) = \Lambda_{\alpha_1,\alpha_2}(x) + \Lambda_{\alpha_2,\alpha_0}(x), \tag{2.35}$$

The relative entropy is the expected value of the log-ratio, with respect to the distribution "in the numerator", for exponential families is

$$\begin{aligned} D(P_{\alpha_1} || P_{\alpha_0}) &= \mathbb{E}_{\alpha_1}(\Lambda_{\alpha_1,\alpha_0}(X)) \\ &= (\alpha_1 - \alpha_0) \cdot \mathbb{E}_{\alpha_1}(f(X)) - (F(\alpha_1) - F(\alpha_0)). \end{aligned} \tag{2.36}$$

The decomposition 2.35 implies

$$D(P_{\alpha_1}||P_{\alpha_0}) = D(P_{\alpha_1}||P_{\alpha_2}) + D(P_{\alpha_2}||P_{\alpha_0}). \tag{2.37}$$

Note that for general distributions P and Q is not even guaranteed that the divergence $D(P||Q)$ respects the triangular inequality. So the properties of the relative entropy tend to be optimal in the exponential family.

2.4 Geometry of Exponential Families

The exponential family is often used because it has optimal mathematical properties. Most of them arise because of the structure of the *log-space* of densities. In particular, the family depends on x only through the sufficient statistics $f(x)$, so the space of log-densities is actually a *vector space* of sufficient statistics, denoted with $\mathcal{F}$. For absolutely continuous distributions, the sufficient statistics are continuous functions of a Lebesgue space, so the vector space is infinite dimensional, whereas for the discrete case, $\mathcal{F}$ is m dimensional. This section is inspired by Jørgensen and Labouriau (2012), and the results in linear algebra needed for the discussion are taken from Cailotto (2004).

Affine Geometry of Finite Dimensional Families

Linear transformations do not change the structure of the vector space. Denote with $\varphi_{\mathbf{B}} : \mathcal{F} \to \mathcal{F}$ the linear function defined as $\varphi_{\mathbf{B}}(f(x)) = \mathbf{B}f(x)$, where $\mathbf{B}$ is a $m \times m$ dimensional matrix. If $\mathbf{B}$ is nonsingular, $\varphi_{\mathbf{B}}$ is bijective, corresponds to a change of basis in $\mathcal{F}$ and will be called *isomorphism*. The parameters link the vector space of sufficient statistics and the distributions in the family. A fixed $x_0 \in \mathcal{X}$ defines a reference point $f(x_0)$ in the vector space. For simplicity it can be assumed that $f(x_0) = 0$ (see Jørgensen and Labouriau (2012), theorem 1.3 for more details). The parameters are the coordinates of the sufficient statistics $f(x)$ with respect to the reference point $f(x_0) = 0$. Therefore, if the vectors in $\mathcal{F}$ are transformed by $\varphi_{\mathbf{B}}$, the link between the distribution in the family and the points in the vector space remains the same when the new parameter is $\beta(\alpha) = \mathbf{B}^{-1}\alpha$.

The interest is when the linear function is not a isomorphisms, as we can use sufficient statistics in a lower dimensional vector space to approximate the complete model. In particular, consider the family

$$Q_\theta(x) = e^{\theta \cdot w(x) - W(\theta)}, \tag{2.38}$$

where $w(x) = \mathbf{W}f(x)$, $\mathbf{W}$ is a $t \times m$ dimensional matrix with $t \leq m$. $\mathbf{W}$ induces the linear function $\varphi_{\mathbf{W}} : \mathcal{F} \to \mathcal{W}$, where $\mathcal{W}$ is a linear subspace of $\mathcal{F}$. If $\text{rank}(\mathbf{W}) = t$, $\varphi_{\mathbf{W}}$ is surjective and $\dim(\mathcal{W}) = t$.

If P_α is approximated by Q_θ, the information loss corresponds to the *null space* of $\mathbf{W}$, defined as the space of vectors that $\varphi_{\mathbf{W}}$ sends to 0. This space is denoted as $\mathcal{U}$, it is orthogonal to $\mathcal{W}$ ($\mathcal{F} = \mathcal{W} \oplus \mathcal{U}$) and its elements are

$u(x) = \mathbf{U}f(x)$, where $\mathbf{U}$ is $(m-t) \times m$ dimensional such that $\mathbf{W}\mathbf{U}^\top = \mathbf{0}$. Then the residual of the approximation of P_α with Q_θ is

$$R_\gamma(x) = e^{\gamma \cdot u(x) - U(\gamma)}, \tag{2.39}$$

which is a family of distribution with sufficient statistics in $\mathcal{U}$.

Note that if the models P_α, Q_θ and R_γ are minimal, their sufficient statistics belong to a m, t and $m-t$ dimensional vector space respectively. For identificability the link between the vector space and the distribution has to be one to one, so it's assumed that the components of α, θ and γ are not equal. Therefore, as P_α is equivalent to every model in which the sufficient statistics are transformed by $\mathbf{B}f(x)$, when $\mathbf{B}$ nonsingular. For identificability, assume that the components of $\mathbf{B}^\top\alpha$ are different. Q_θ is equivalent to all models with sufficient statistics $\mathbf{B}_w w(x)$, with $\mathbf{B}_w$ m-dimensional nonsingular, and $\mathbf{B}_w^\top\theta$ has different components. Analogously R_γ is equivalent to all family with sufficient statistics $\mathbf{B}_u u(x)$ and parameter $\mathbf{B}_u^\top\gamma$, where $\mathbf{B}_u$ is an $(m-k) \times (m-k)$ nonsingular matrix. $\mathbf{W}\mathbf{U}^\top = \mathbf{0}$ implies that

$$\mathbf{B}_w\mathbf{W} \cdot (\mathbf{B}_u\mathbf{U})^\top = \mathbf{0}_{t,t-m}. \tag{2.40}$$

So, the decomposition $\mathcal{F} = \mathcal{W} \oplus \mathcal{U}$ is maintained after the transformation of $w(x)$ and $u(x)$.

Infinite Dimensional Exponential Families

The algebraic results introduced so far, works only when the configuration space is finite dimensional. In fact, in this case every *endomorphisms* (invertible linear functions from a space to itself) are bijective, and so they are isomorphisms. Instead when the space of sufficient statistics is infinite dimensional, there are endomorphisms which are injective but not surjective or vice versa (Cailotto (2004), chapter 2).

In an infinite dimensional space we can define a *infinite dimensional exponential family* of distributions with densities

$$dP_{f,h}(x) = e^{f(x)-F(f)}d\nu_h(x), \tag{2.41}$$

where $d\nu_h(x) = h(x)d\nu(x)$ is the base measure of the space, ν is the counting or the Lebesgue measure if $\mathcal{X}$ is discrete or continuous respectively. The measurable function $f : \mathcal{X} \to \mathbb{R}$ serves as parameter in an infinite dimensional space of real functions $\mathcal{F}$, which is a linear space as $af + bg \in \mathcal{F}$, when $f, g \in \mathcal{F}$ and $a, b \in \mathbb{R}$.

A t-dimensional approximation of $P_{f,h}$ is

$$dQ_{w,\theta,h;t}(x) = e^{\sum_{i=1}^t \theta_i w_i(x) - W_t(\theta)}d\nu_h(x), \tag{2.42}$$

where $\sum_{i=1}^t \theta_i w_i(x) < \infty$ for all $x \in \mathcal{X}$, t can be both finite or infinite, and fix the precision of the approximation. The set of linearly independent functions

$w_1, ..., w_t$ is a basis of $\mathcal{W}_t$, some examples are polynomials, splines or trigonometric functions. In some cases we have that

$$\lim_{t\to\infty} \mathcal{W}_t = \mathcal{F}, \tag{2.43}$$

and the approximation error converges to 0. In fact, in this case exists an infinite dimensional vector θ^* such that $\sum_{i=1}^{t} \theta_i^* w_i(x) \xrightarrow{t} f(x)$, $W_t(\theta^*) \xrightarrow{t} F(f)$ and

$$dQ_{w,\theta^*,h;t}(x) \xrightarrow{t\to\infty} dP_{f,h}(x), \tag{2.44}$$

for all $f \in \mathcal{F}$.

Not all density functions can be written as 2.41, however in some cases it is still possible having no approximation error, but only when t is infinite dimensional. Moreover, in some cases it's possible derive the convergence rate $D(P||Q_t^*)$ for finite t, where P is the real distribution with unknown density, and $Q_t^* = Q_{w,\theta^*,h;t}$ is the finite dimensional exponential family generated by $w_1, ..., w_t$. Q^* is called *information projection* and

$$\theta^* \in \Theta_t = \{(\theta_1, ..., \theta_t) : \ \theta_i \in \mathbb{R}\}. \tag{2.45}$$

The theory of this approximation method will be developed in the next section, in particular the information projection does not always exist.

If P can not be written as 2.41, also the base measure h have to be chosen. When the base measure of the real density ν_h is unknown, it can be replaced with $\nu_{\tilde{h}}$ only if

$$\text{supp}(\nu_{\tilde{h}}) = \text{supp}(\nu_h). \tag{2.46}$$

Intuitively, the base measure should approximate as much as possible ν_h, however the base measure can not be estimated (remaining in the context of exponential families). If the sufficient statistics w_i are infinitely differentiable, like polynomials for example, the function $\tilde{h}$ fixes the smoothness of the family, in term of how many times Q_m is differentiable. The principle of maximum entropy can be used also when there is no information available, if $\mathcal{X}$ is bounded, the solution is the uniform reference measure:

$$\text{argmax}_{\tilde{h}} H(Q_{\tilde{h}}) = 1. \tag{2.47}$$

This choice can be used also when P, the real density can not be written as 2.41. The only information needed of P is its support, as $\tilde{h}$ fixes the support of all distributions in the approximating family. In the most general form of infinite dimensional exponential families the sufficient statistic $f(x)$ can be modelled directly in the function space without using basis of polynomials, considering f in a *reproducing kernel Hilbert space*, see Canu and Smola (2006). This theory is famous in statistics especially for the *support vector machine* or more generally in *nonparametric Bayesian statistics*.

Curved Families

So far, the approximating family has been derived using a lower dimensional space of sufficient statistics, both in the finite or infinite cases. In the first one has been used a linear transformation of low rank, in the second case the space of all real functions has been reduced with a finite or infinite dimensional space with continuous basis functions. A completely different approach is reducing the dimension of the parametric space. If the problem is discrete, this method is particularly useful when a minimal exponential family is too large to describe the system under study, because most combinations of the parameters lead to a "pathological" distributions, called *degenerate*. Also, the method can be useful to reduce the correlation between the sufficient statistics.

A *curved exponential family* $P_{\alpha(\eta)}(x)$, where $\alpha : \mathbb{R}^m \to A_\eta \subset \mathbb{R}^m$, and A_η is curved subspace of $\mathbb{R}^m$, parametrized by the t-dimensional parameter η. For different values of η, the densities in the family are

$$dP_{\alpha(\eta)}(x) = e^{\sum_{i=1}^m \alpha_i(\eta) f(x) - \widetilde{F}(\eta)} d\nu_h(x), \tag{2.48}$$

where $\widetilde{F}(\eta) = F(\alpha(\eta))$. The log space of this family is not an affine space, as A_η does not contain a m-dimensional neighborhood. For more information on curved exponential families see Efron et al. (1978).

2.5 Minimum Information Projection

In the previous section has been shown how the geometric properties of the exponential family allow to obtain an approximation of a "complete" family P_α with a "low rank" approximating family Q_θ. This approach works only when the log-space of P_α is a finite dimensional vector space, or if the true density can be written as an infinite dimensional exponential family. However not all distributions belong to this class. The method introduced in this section finds the distribution in a family closest to a convex set of distributions, specified by some restrictions.

Denote with $\mathcal{P}$ the set of distributions that describe the system under study. Fix $Q_0 \in \mathcal{P}$, and consider the *information sphere* of radius ρ of distributions close to Q_0 in information sense:

$$S(Q_0, \rho) = \{P \in \mathcal{P} : D(P||Q_0) \leq \rho\}. \tag{2.49}$$

Let $\mathcal{C} \subset \mathcal{P}$ a convex set of probability distributions such that $\mathcal{C} \cap S(Q_0, \infty) \neq \emptyset$. If exist $Q^* \in \mathcal{C}$ such that

$$Q^* = \underset{P \in \mathcal{C}}{\operatorname{argmin}}\ D(P||Q_0), \tag{2.50}$$

then Q^* is unique (by convexity of $\mathcal{C}$) and it's called *information projection* of Q_0 in $\mathcal{C}$.

The existence of Q^* will be described later. First note that when such Q^* exists, the relative entropy respects the triangular inequality

$$D(P||Q_0) \geq D(P||Q^*) + D(Q^*||Q_0), \tag{2.51}$$

for all $P \in \mathcal{C}$. Moreover if for all $P \in \mathcal{C}$, exists $P' \in \mathcal{C}$ such that $Q^* = \alpha P + (1-\alpha)P'$, then

$$D(P||Q_0) = D(P||Q^*) + D(Q^*||Q_0), \tag{2.52}$$

for all $P \in \mathcal{C}$ (Csiszár (1975), theorem 2.2). Equation 2.51 assures that the information loss on approximating P with Q_0 is greater than the information loss using Q^* as an intermediate point. For specific sets $\mathcal{C}$ the stronger result in equation 2.52 allows a decomposition of the approximation loss similar to the one in equation 2.37 for exponential families.

The case which is most interesting to us is when the set $\mathcal{C}$ is defined by linear constraints

$$\mathbb{E}_P(f_i(X)) = \int f_i(x)dP(x) = \mu_i < \infty, \tag{2.53}$$

for $i = 1, ..., m < \infty$, and $P \in \mathcal{P}$ is a fixed (real) distribution of the experiment. Denote with $\mathcal{P}_{\mu_i} \subset \mathcal{P}$ the set

$$\mathcal{P}_{\mu_i} = \left\{ \widetilde{P} \in \mathcal{P} : \ \mathbb{E}_{\widetilde{P}} f_i(x) = \mu_i = \mathbb{E}_P f_i(x) \right\}. \tag{2.54}$$

Then

$$\mathcal{C} = \mathcal{P}_\mu = \bigcap_{k=1}^{m} \mathcal{P}_{\mu_i}. \tag{2.55}$$

is the set of probability distributions with same expected sufficient statistics as P.

Theorem 3.1 in Csiszár (1975) gives the necessary condition that the information projection of Q_0 in $\mathcal{P}_\mu$, if exists, it has density

$$dQ^*(x) \propto \exp\left(\sum_{k=1}^{m} \theta_i^* w_i(x) \right) dQ_0(x). \tag{2.56}$$

But, when $\theta = 0$, Q_θ is the uniform distribution, so the information projection Q^* has probability density function

$$q_m^*(x) = \exp(\theta^* \cdot w(x) - W_m(\theta^*)). \tag{2.57}$$

Then, for the same theorem, Q^* allows the decomposition 2.52. Combining this decomposition with 2.37, we have that

$$D(P||Q_0) = D(P||Q^*) + D(Q^*||Q_{\theta_0}), \tag{2.58}$$

for every fixed $\theta_0 \in \Theta$. Therefore Q^* can be seen as the projection of the whole

family Q_θ. If it exists is unique and

$$Q^* = Q_{\theta^*} = \mathcal{P}_\mu \cap Q_\theta. \tag{2.59}$$

If a sample from X is available, let

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} f_j(x_i), \tag{2.60}$$

for $j = 1, ..., m$, $\hat{\theta}$ is the solution of

$$\mathbb{E}_{Q_\theta}(f(X)) = \hat{\mu}. \tag{2.61}$$

The distribution $\widehat{Q} = Q_{\hat{\theta}}$ is the estimator of Q^* that approximates P. By 2.58 the information lost from P to $\widehat{Q}$ is the sum of the *approximation error* $D(P||Q^*)$ and the *estimation error* $D(Q^*||\widehat{Q})$.

Parametric Density Estimation

This approach has been used in the problem of estimating continuous distribution $P \in \mathcal{P}$ in a bounded domain, say $\mathcal{X} = [-1, 1]$. In Crain (1977) the assumption is that $\mathcal{P}$ is an infinite dimensional exponential family as 2.41, and the approximating density is its t-dimensional approximation, using orthogonal polynomials as sufficient statistics. Then, Barron and Sheu (1991) generalize the set $\mathcal{P}$ and shows that the information projection converges to an infinite dimensional exponential family when the number of sufficient statistics goes to infinity. Their theory shows that distributions in some spaces $\mathcal{P}$ can be approximated with arbitrary accuracy by the information projection when $m \to \infty$. The accuracy depends on the rate of convergence, which is derived when the sufficient statistics of the approximating family are orthonormal.

If the density is strictly positive in its domain than it can be written as

$$p(x) \propto \exp\left(\sum_{k=0}^{\infty} \theta_k \varphi_k(x)\right), \tag{2.62}$$

where $x \in [-1, 1]$ and $\varphi_k(x)$ is the *Legendre polynomial* of degree k, these polynomials are orthogonal with respect to the L_2-norm in $[-1, 1]$:

$$\int_{-1}^{1} \varphi_i(x)\varphi_j(x)dx \propto \delta_{ij}, \quad 0 \le i, j < \infty. \tag{2.63}$$

Then, using the fact that $\int_{-1}^{1} p(x)dx = 1$, the density can be written as

$$p(x) = \exp\left(\sum_{k=0}^{\infty} \theta_k \varphi_k(x) - F_\infty(\theta)\right), \tag{2.64}$$

and can be approximated with

$$p_m^*(x) = \exp\left(\sum_{k=0}^{m} \theta_k \varphi_k(x) - F_m(\theta)\right), \tag{2.65}$$

using m Legendre polynomials as sufficient statistics.

In this context, it is explicit how the approximation works. The approximating density is the information projection of p_m, the exponential family with Legendre polynomials as sufficient statistics. The knowledge of all infinite expected values of these polynomials corresponds to all information in p. Therefore the exponential family p_m is *dense* in the set of bounded continuous densities in $[-1, 1]$. There are generalizations of this result to more difficult cases, such as densities in unbounded domains, however for each problem, there are assumptions in the form of p. These assumptions reflect some properties on the real density are generally in the *tails* of the distribution, and the *smoothness* of $\log(p)$, in term of how many time is differentiable with respect to x.

Barron and Sheu (1991) develop further this approach considering the cases in which the log-density is approximated by a spline or a linear combination of orthogonal polynomials or trigonometric series. They derive the rate of convergence of the approximation in term of the Kullbach-Leibler divergence in different set up of smoothness assumptions and in all three cases of sufficient statistics. However, for the polynomial case, which is the only one exposed here, the rate of convergent is derived only when the polynomials are standardized to be *orthonormal* as this condition leads to an optimal rate. Let $\Lambda_m(x) = \log p(x) - \log p_m^*(x)$ the *log-ratio* between the real and approximating density. The rate of convergence is

$$D(P||P_m^*) = O(||\Lambda_m(x)||_2^2). \tag{2.66}$$

This rate of convergence is valid when $||\Lambda_m(x)||_\infty$ is bounded, exist A_m such that $||\log p_m||_\infty \leq A_m||\log p_m||_2$ for all densities p_m of distributions in the family P_m, and $A_m||\Lambda_m(x)||_2 \xrightarrow{m} 0$. If a n dimensional sample from P is available, the estimation error converges to 0 in probability if $A_m\sqrt{m/n}$, as

$$D(P_m^*, \widehat{P}_{n,m}) \leq O_{pr}\left(\frac{m}{n}\right), \tag{2.67}$$

where $\widehat{P}_{m,n} \in P_m$ is the distribution with parameter $\hat{\theta}$ that corresponds to the maximum likelihood estimator, which is the value that solves equation 2.61.

The number of sufficient statistics in the approximating family can be chosen "automatically", specifying $m = m(n)$. For example, consider the *Sobolev space of functions* W_2^r. This space contains the functions such that their $(r-1)$-th derivative $f^{(r-1)}$ is absolutely continuous and $\int (f^{(r)}(x))^2 dx < \infty$. If the

logarithm of the real density $f(x) = \log p(x)$ belong to W_2^r then

$$D(p||\hat{p}_{n,m}) = O_{pr}\left(\left(\frac{1}{m}\right)^{2r} + \frac{m}{n}\right), \tag{2.68}$$

and if $m(n) \propto n^{1/(2r+1)}$, then

$$D(p||\hat{p}_{n,m}) = O_{pr}\left(n^{-2r/(2r+1)}\right), \tag{2.69}$$

which is the optimal *minimax rate* for distributions with log-densities in W_2^r, see Barron and Sheu (1991) for more information.

Chapter 3

Random Networks

Introduction

The term *network* is widely used and somehow vague in what represents. In fact numerous systems from very different fields can be represented using a network. A very intuitive definition is a collection of connected objects, which are called *nodes*, *actors* or *vertices*, the connections between them are *ties*, *edges* or *links*. An example from chemistry is a molecule, which can be seen as collection of atoms (nodes) linked by chemical bonds which are the edges. Completely different examples are the airline network, where the airports are connected by flight routes, or the brain, where the vertices are neurons and the edges are synapses. The interest is typically in the structure of these networks, which in the previous examples influences the function of the molecule, the air traffic and the capacity of the brain. Of course the systems described by a network in the examples are very different from each other. Nevertheless the mathematical formalism allows to define and analyse properties of the network regarding the source of the data. In mathematics networks are called more frequently *graphs*. In this thesis the focus will be on *random networks*, in which the connections between vertices are random according to a probability distribution. Statistically, this formalism allows to analyse the network as an object, rather then model each probability of a connection.

In this chapter is introduced the mathematical formalism of networks. In particular is exposed the information that we can gather from a graph, which is in term of descriptive statistics. Many of them are based on the counts of particular (usually small) pattern of ties in the network, these structures are called subgraphs. The emphasis is in the ones involved with the eigenvalue distribution of the network, which will be shown that it contains all the information on the network, when it is homogeneous. The graph that represents the system under study is formed by ties and the information we can gather is based in their placement. If two vertices are connected by a tie they are *adjacent*, this chapter starts with a discussion of which relations between vertices the ties can represent.

3.1 Adjacency Relations

Like stated previously in the introduction, a network is a collection of nodes and edges. For random graphs, the nodes are considered fixed and the connections between them are random. The network is so a form of *relational data*, "the information contained in the object can not be reduced to the characteristics of the nodes" (Handcock et al. (2008)). The observed ties (i.e. the observed graph), in the context of random networks, can be considered as an event in the sample space of possible edge sets. The relations between actors can be of various types, each of them leads to a different class of networks.

The first distinction depends if the edges are *binary* (the relation between individuals can be present or not) or *weighted* (the relations have weights, depending on their strength). These types of edges lead to the classes of *unweighted* and *weighted networks*. In the first case "only" the structure of the network is the object of the analysis, in the second case the network includes all possible edges (eventually with weight 0 if there is no connection) and the object of the analysis is the *weight distribution*. The last main distinction is in *directed* and *undirected networks* if the relations are *unilateral* or *bilateral* respectively. For directed networks the edges are represented using arrows and sometimes the relations can be interpreted causally, for undirected graphs the connections are represented by lines between nodes. Often there are restrictions on edges that connect vertices to themselves (called *loops* or *self-loops*), or if multiple connections between two nodes are allowed or not. The mathematics involved in these classes of network is often very different, in this thesis the focus will be only on undirected networks without self-loops and multiple edges between, these are called *simple networks*.

In a simple graph, the edges can be codified as binary variables. The network is so represented by the *adjacency matrix* $\mathbf{A} = \mathbf{A}(G)$ defined as $\mathbf{A}_{ij} = 1$ if $ij \in E(G)$, 0 elsewhere. On the other way around, every symmetric n-dimensional matrix, with null entries in the diagonal, and either 0 or 1 outside the diagonal is the adjacency matrix of a simple network. Therefore this matrix is a proper representation of a network and thus it has all information of the graph. Functions of the adjacency matrix can be interpreted as functions of the represented network.

Many properties of the graph depend on the structures present in the network. For example the triangles represent the construct in which if the actors x and y are friends, and also y and z, then the probability that x and z are friends too is usually higher than the probability of friendship between random actors. Many other structures can represent non-trivial patterns in the ties. These are called *subgraphs* (or *subnetworks*) and they are important especially when they are low dimensional. More formally if $G = (V, E)$ is a simple graph, a subgraph is $G^* = (V^*, E^*)$ such that $V^* \subseteq V$ and $E^* \subseteq E$. By counting of small subgraphs we can measure *local properties* of the network, useful in the description of the system when ties which are close to each other are conditionally dependent.

Some important subgraphs are:

- k-dimensional *walk* (or k-walk): sequence of actors $v_1, v_2, ..., v_{k+1}$, $v_i \in V(G)$. If $v_1 = v_{k+1}$ is a *close walk*, the total number of k-dimensional close walks is denoted as $W_k = W_k(G)$.
- *cycle* of length k: close walk $v_1, v_2, ..., v_{k+1} = v_1$ such that $v_i \neq v_j$ for all $i \neq j$. Is denoted as C_k, C_3 is a triangle, the total number of close walks in C_3 is $M_3(C_3) = 6$.
- *k-star*: $k+1$ dimensional graph with one node called root connected to all other k vertices, which are not connected between themselves, therefore they have degree 1. The k-star is denoted as S_k.
- *k-triangle*: union of k triangles with a common edge between all of them.
- j-dimensional *complete graph*: set of vertices in which every combination of actors is connected. This graph is denoted with K_j.
- *clique*, maximal complete subgraph of a network, i.e. set of actors $v_1, ..., v_k$, $v_i \in V(G)$ such that the subgraph of G spanned by these vertices is complete. The set of k-dimensional cliques of G is denoted with $\widetilde{\mathscr{C}_k} = \widetilde{\mathscr{C}_k}(G)$. The same set without multiple occurrences is denoted as $\mathscr{C}_k$.

If the vertices are unlabelled, it is not important which vertices connect an edge. However in this case every structure of ties can be represented by different adjacency matrices. The way to introduce equivalence classes of matrices that represent the same network is done through the notion of *isomorphism*. This is the bijective map

$$\alpha : V(G) \to V(H) \quad \text{s.t. } xy \in E(G) \iff \alpha(x)\alpha(y) \in E(H), \tag{3.1}$$

the isomorphism α preserves adjacency (edges mapped to edges) and non-adjacency (non-edges mapped to non-edges).

If such function α exist, then the graphs G and H represent the same relations between actors. Note that the number of vertices $|V(G)|$ has to be equal to $|V(H)|$, more generally also the number of edges, triangles and all other structures are the same in both graphs. If the network is unlabelled a good model has to be invariant with respect to the permutation isomorphism. Other models include the possibility that the vertices belong to different classes. In this case the invariance can be with respect to permutations between vertices of the same class.

3.2 Descriptive Statistics

Our interest is on properties of the system that the network represents, whether it is biological, social and so on. Some of its properties are related with mathematical quantities associated with the graph that represent the system. If

network is random the ties have a joint probability distribution in the space of possible adjacency matrices. So, we can gather information contained in the probability distribution of the network using sufficient statistics of the adjacency matrix. This section introduces many quantities related to properties of a network. They are usually one or n dimensional, some of them describe the network focusing on local properties, others are more influenced by the global structure of the network.

Given a graph G with n vertices, the simplest information that we have on the distribution of the edges is the *average degree*. This is

$$\bar{d} = \frac{2|E(G)|}{n} = \frac{1}{n}\mathbf{1}^T\mathbf{A}\mathbf{1}, \tag{3.2}$$

$\bar{d}$ is really important in modelling the network because the models perform very differently depending on the relation between $\bar{d}$ and n. Let's define G_n, $n = 1, ...$ the sequence of growing graphs with n vertices. A random network is called *sparse* if $n\bar{d} = 2|E(G_n)| = O(n)$, and *dense* if $|E(G_n)| = O(n^2)$. Even though only one n-dimensional network is observed, most of the properties of a random model depend on the sequence G_n, so a model may perform well only for a certain range of total number of connections. Almost all real world networks are sparse, especially when they are large. Another important statistics are the *number of connected components* and *number of isolates nodes*. Some descriptive statistics and models are useful only when the graph is connected, for example when they involve the use of the eigenvectors. Also, we may want restrictions on the information contained in different components, for example forcing conditional independence between actors (or ties) if they belong to different components.

In many real world networks, the number of triangles observed is often higher than expected for a random model. This effect is measured by the *cluster coefficient*

$$C = \frac{3|C_3|}{|S_2|}, \tag{3.3}$$

the denominator is the number of two stars, C measure how many of them are "closed" to form a triangle (the constant in the numerator is because every triangle contains three 2-stars). Many other subgraphs can be measured to compare the observed network with a random model. These subgraphs are called *motifs*, and with every network G is associated the *significance profile* $SP(G) = (SP_1, ..., SP_m)$

$$SP_i(G) = \frac{Z_i(G)}{\sum_j Z_j(G)^2}, \quad Z_i(G) = \frac{|\mathscr{S}_i(G)| - \mu_{\mathscr{S}_i}}{\sigma_{\mathscr{S}_i}}, \tag{3.4}$$

where $|\mathscr{S}_i(G)|$ is the count of the motif $\mathscr{S}_i$ in G, $\mu_{\mathscr{S}_i}$ and $\sigma_{\mathscr{S}_i}$ are the expected mean and standard deviation of the subgraph count in the random model used for comparison (Estrada (2012), chapter 4). The occurrence of motifs in various networks (gene regulation, neurons, food webs, electronic circuits and World

Wide Web) has been studied in Milo et al. (2002).

To compute the significance profile, the subnetworks $\mathscr{S}_1, ..., \mathscr{S}_m$ have to be counted. The subgraphs counts are functions of the adjacency matrix, for some of them this function is explicit. Important subgraphs that describe the local topology of the network are the closed walks. In Van Mieghem (2010) is shown that the number of k step walks from the i-th to the j-th actor is $(\mathbf{A}^k)_{ij}$, so the i-th diagonal elements of the powers of the adjacency matrix are the number of closed walks that start and end in the i-th actor. Since the sum of the diagonal elements of a adjacency matrix is equal to the sum of its eigenvalues, the total number of k step closed walks is

$$W_k = \sum_{i=1}^{n} \lambda_i^k = \mathrm{tr}(\mathbf{A}^k), \tag{3.5}$$

where $\lambda = (\lambda_1, ..., \lambda_n)$ is the *eigenvalue sequence* of the graph. The eigenvalues are almost only used in undirected network, because the adjacency matrix is symmetric, therefore they are real numbers.

There is an explicit relation between the counts of particular subgraphs and the closed walks W_k, the first term are

$$\begin{aligned} W_2 &= 2|E| \\ W_3 &= 6|C_3| \\ W_4 &= 2|E| + 4|S_2| + 8|C_4| \\ W_5 &= 30|C_3| + 10|\triangleright\!\!-| + 10|C_5|. \end{aligned} \tag{3.6}$$

For $k \geq 6$ the relation become more difficult. Up to $k \leq 5$ the odd walks are influenced only by "odd" sugbraphs (structures that does not involve triangles), while even walks are influenced only by "even" subgraphs. Instead W_6 is influenced also by the number of 2-triangles, essentially an odd structure. However, the small subgraphs tend do be important in W_k when k is big, except for E, which influences all even subgraphs by the same factor $2|E|$. In a small subgraph spanned by a long closed walk, there are many possibilities in which the walks can cross multiple times some edges, while the walk remain close, so W_k, also when k is big, are heavily affected by counts of small graph. This explain intuitively why the closed walks are important in describing the local topology.

It is possible to combine walks of different length in a one dimensional statistic. The following linear combination can be used

$$\sum_{k=0}^{n} c_k W_k(G) \leq \sum_{k=0}^{n} c_k W_k(K_n) = \sum_{k=0}^{n} c_k((n-1)^k + (n-1) \cdot (-1)^k), \tag{3.7}$$

where c_k is a sequence of constants that guarantee that the sum converges for every possible graph, i.e. it converges for the complete graph K_n. The main one-dimensional index defined in this way is the *Estrada index* of the graph G

(Estrada (2012), chapter 5), which is

$$EE(G) = \sum_{k=0}^{\infty} \frac{W_k}{k!} = \sum_{k=0}^{\infty} \frac{1}{k!} \mathrm{tr}(\mathbf{A}^k) = \mathrm{tr}\left(\sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!} \right) = \mathrm{tr}(e^{\mathbf{A}}), \tag{3.8}$$

where $\mathbf{A}$ is the adjacency matrix of G.

The index is bounded by

$$n = EE(0_n) \leq EE(G) \leq EE(K_n) < e^{n-1} + \frac{n-1}{e}. \tag{3.9}$$

The index uses $c_k = \frac{1}{k!}$ as standardization constants, this can be generalized to $c_k = \frac{\beta}{k!}$, $\beta > 0$ and is called *parametrized Estrada index*

$$EE(G, \beta) = \mathrm{tr}(e^{\beta \mathbf{A}}). \tag{3.10}$$

This statistics gives a statistical mechanics interpretation of the network. The system represented by the graph can bee seen as immersed in a *heat bath* at *temperature* T, that is an external situation that affect equally all possible ties. The parameter $\beta = \frac{1}{k_B T}$ is the *inverse temperature* (k_B is the *Boltzmann constant*). Therefore the inverse temperature weight all possible edges of the network. In the limit $\beta \to 0$, the network converges to the empty graph, and the vertices can be seen as particle of a gas, while when $\beta \to \infty$ the network is a complete graph in which all edges have infinite weight, this configuration reminds a solid state. In the last case

$$\frac{EE(G, \beta)}{e^{\beta \lambda_1}} \xrightarrow{\beta \to \infty} 1, \tag{3.11}$$

so only the biggest eigenvalue is relevant in the structure. Other results in statistical mechanics are in Estrada (2012), chapter 5 and in Estrada and Hatano (2007).

Other two important indexes are derived from the decomposition

$$\begin{aligned} EE(G) = \mathrm{tr}(e^{\mathbf{A}}) &= \sum_{k \text{ odd}} \frac{W_k}{k!} + \sum_{k \text{ even}} \frac{W_k}{k!} = \\ &= \mathrm{tr}(\sinh(\mathbf{A})) + \mathrm{tr}(\cosh(\mathbf{A})) = EE_{\mathrm{odd}}(G) + EE_{\mathrm{even}}(G). \end{aligned} \tag{3.12}$$

A *bipartite graph* is a network in which the vertices can be divided in two communities such that there are no edges between vertices in the same community (a *tree* is an example of bipartite graph). In this case there are no odds walks in the network, therefore $EE_{odd}(G) = 0$. If the index is close to 0, or in general much smaller than the even counterpart the network is *almost bipartite*, so relations between people in the same community are possible but much more rare than relations between people in different communities.

Centrality Measures

A network is an $n(n-1)/2$ dimensional binary object, we can summarize some of the information contained in it with a sequence of n numbers that hopefully describes properties of the network. The first way introduces measure the importance of the n actors based on the disposition of the ties. These are called *centrality measures*, the simplest one is the *degree sequence* or *degree centrality* d defined as

$$d = (d_1, ..., d_n), \quad d_i = |N(i)|, \tag{3.13}$$

where $N(i)$ is the neighborhood of the i-th vertex, i.e. how many friends the i-th actor has. The statistical properties of this sequence are in the *degree distribution*:

$$\tilde{d}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{d_i}(x), \tag{3.14}$$

where $\delta_a(x)$ is the *Dirac's delta function*, defined as $\delta_a(x) = 1$ if $x = a$, 0 elsewhere. Thus, $\tilde{d}(x)$ is the proportion of vertices with degree x.

The degree sequence is a *first order centrality measure*, only neighbours affect the importance of a vertex. Consider a random walk in the graph G (assume G connected) that starts from a fixed vertex and in each point move to a vertex in the neighborhood with constant probability. If the walk is long enough, the starting vertex is not important as it can explore the whole network, so its distribution became independent on the starting position. Every d_i is proportional to the time spent in the i-th vertex by infinite dimensional random walk defined as before. This is called sometimes *random walk centrality measure* and is essentially equivalent of the degree centrality d.

The degree distribution can be relevant on explaining pattern of ties when actors with many friends are more important in the global structure of the network. The degree distribution is the statistic that has been studied the most in analysis of random networks. The subgraphs associated with this distribution are the *k-stars* S_k. Each S_k contains $\binom{k}{j}$ j-stars for all $j < k$ (Frank and Strauss (1986)), so there is the one-to-one relation

$$|S_k| = \sum_{j \geq k} \binom{j}{k} \tilde{d}(j). \tag{3.15}$$

The last equation shows the explicit relation between a network statistic, the degree distribution, and the count of particular subgraphs, the k-stars.

On the other extreme, a centrality measure that characterize the global structure of the network is the *eigenvector centrality* $u = (u_1, ..., u_n)$ where u is the eigenvector of the adjacency matrix associated with the biggest eigenvalue. This centrality is defined only when the network is connected, in this case the elements of u are all positive. Every element u_i is equal to the ratio by all infinite dimensional open walks that start in i, with the total number of infinite dimensional open walks. If the open walk is very long, after some point he will has crossed all vertices, so he lost the information in the neighborhood of

the starting point, i.e. the local structure around the i-th actor. That's why long open walks characterize the global structure of the network, whereas closed walks are associated with the local structure. The values u_i give the relative importance of the vertices in the global structure of the network and u is a *global centrality measure*.

A compromise between the two centrality measures previously defined is an index influenced by the local behaviour of the network, however considering more structure than d, which can be considered as a first order centrality measure. This index is associated with the *eigenvalue sequence* of the adjacency matrix $\lambda = (\lambda_1, ..., \lambda_n)$. Like for the degree case, more information is in the *eigenvalue distribution*

$$\tilde{\lambda}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i}(x). \tag{3.16}$$

The moments of this distribution are

$$\frac{W_k}{n} = \int x^k d\tilde{\lambda}(x) = \frac{1}{n} \sum_{i=1}^{n} \lambda_i^k = \frac{1}{n} \mathrm{tr}(\mathbf{A}^k) \tag{3.17}$$

$k = 1, 2, ...$ These quantities are important because W_k is the number of closed walks of k-steps in the network represented by the adjacency matrix $\mathbf{A}$.

The indices introduced below can be generalized to some *local centrality measures*, counting close walks that start from all vertices. The *Estrada* (or *subgraph*) *centrality* is

$$\widetilde{EE} = \mathrm{diag}(e^{\mathbf{A}}) = \sum_{k=0}^{\infty} \frac{1}{k!} \mathrm{diag}(\mathbf{A}^k), \tag{3.18}$$

$(\mathbf{A}^k)_{ii}$ is the total number of walks that start and end in node i. There are explicit formulas that link these closed walks to the subgraphs that contains the node i. They are similar to the ones in equation 3.6, however "non-symmetric" subgraphs (like $\rhd\!-$ for example) appear multiple times because the location of i in the subgraph is important (see Estrada (2012), chapter 7). Similar to the one dimensional statistics, the *parametrized*, *odd* and *even Estrada centrality* can be defined considering the matrix $\beta\mathbf{A}$ with inverse temperature $\beta > 0$ for the first, and only the odd, or even, powers of $\mathbf{A}$ for the last ones.

3.3 Models for Random Networks

In the previous section have been introduced various ways to summarize the information in a network. If the graph is random, these statistics are heavily influenced by the joint probability distribution of the edge set, which is a *network model*. The definition of a model assigns probabilities to all possible networks, i.e. to all possible combinations of edges.

Formally, the model has to specify how the observed graph has been gener-

ated. The first way to do so is fixing a *generator mechanism*. Various algorithms for generating networks are proposed in this section, some of them mimic how the network evolves. For example the generator mechanisms can be a process in which the network grows until reach a configuration with the same number of actors than the observed one, or the evolution of the network based on local processes between the actors. If the model represent accurately the observed network, it's possible use it for prediction or to test hypothesis on the generator mechanism.

The generator mechanism can be interpreted as an algorithm to generate probability distributions over the space of simple networks. The simplest algorithm is due to Erdös and Rényi (1959). In this model all the ties between n actors are independent with constant probability p, which is the only parameter. It's called *Erdös-Rényi random graph*, despite the simplicity of the mechanism, it presents many non trivial properties (see Newman (2010), chapter 12), one of them is the degree distribution. Consider the n dimensional graph G, let p depends on the size of G as $p = q/(n-1)$, so that q is the expected number of neighbours that has an actor in the network (p and n are linear related so the graph is sparse). The probability that an actor hs d neighbours is

$$p_d = \binom{n-1}{d} p^d (1-p)^{n-1-d} \simeq e^{-q} \frac{q^d}{d!}. \tag{3.19}$$

Asymptotically, when the graph is sparse, the degree distribution is Poisson with mean $q = np$. When this happens the network is connected with high probability. The cluster coefficient is

$$\frac{q}{n-1} \tag{3.20}$$

that goes to 0 when $n \to \infty$, this is one of the problem of this model when it's used to representing real networks.

In the Erdös-Rényi model the probability of having a tie is constant, so is also the expected degree of a vertex. This is always an unrealistic assumption in real world network. The *configuration model* is a generalization for generating graphs uniformly from the ones with a fixed degree sequence. Starting from $d = (d_1, ..., d_n)$, with $\sum_i d_i$ even, the i-th vertex has d_i "pieces of edges" attached to it. Then the vertices are joined accordingly to their degree uniformly in the space of all possible configurations available. Mathematically is simpler allowing the mechanisms to form self loops or multiple edges, however, if the graph is sparse (like in real world networks), the probability to obtain a graph which is not simple vanish with n (Newman (2010), chapter 13).

Most real world network have a *fat tailed* degree distribution. The most important is the *power law distribution* $p_d = d^{-\alpha}/\zeta(\alpha)$ often written in logarithmic scale

$$\log p_d = -\alpha \log d - \log \zeta(\alpha), \tag{3.21}$$

where α is a positive parameter. The moments of order $m \geq \alpha - 1$ diverge.

For most real world network $2 < \alpha < 3$, in such cases the expected degree is defined, however all higher moments diverge, therefore the average degree $\bar{d}$ has infinite variance. In most cases the degree distribution of an observed network fits the power law curve only in the right tail. Therefore, the degrees greater than the parameter d_{min} are power law, and the remaining part of the distribution is modelled in another way, or it is excluded if it's assumed that only the nodes with many neighbours (called *hubs*) are important for the structure of the network.

In some networks the vertices are labelled depending on which community they belong. In this case the straightforward generalization of the Erdös-Rényi random graph is the *stochastic block model*. The ties are generated independently according to the probabilities $p_{01}, ..., p_{0m}$ if both actors belong to the same community, instead, $p_{12}, p_{13}, ..., p_{m-1,m}$ are the probability when the actors belong to different communities. The total number of parameter is $m(m-1)/2 + m$, m is the number of communities. This model it's like a mixture of Erdös-Rényi graphs, so the degree distribution is a mixture of Poisson. Despite so, this model have computational advantages, especially when n is large, and there are many algorithm for estimating the communities, so that they doesn't have to be known a priori. The stochastic block model is invariant under permutations of actors in the same communities. The case with $m = n$ is sometimes called the *Bernoulli model* the tie that connect i and j is generated independently with probability p_{ij}, for $i = 1, ..., n-1$, $j = i+1, ..., n$. If the network is even moderately large, this model is too flexible despite (the usually unrealistic) assumption that the ties are independent.

In the models introduced so far, the joint probability distribution of the ties is derived defining a (stochastic) process that describe the network formation, usually an algorithm that generates graphs. However, this process does not mimic the generator mechanism of the real network, which typically is unknown, and likely too complex to be defined as a mathematical algorithm. Looking the model in a statistical perspective, allows a more direct relation between the distribution of the edge set and the one of the statistics used to summarize the information in the network. Moreover these statistics are counts of particular subgraphs that have relations with properties of the network. Some choices of subgraphs induce a link between the joint probability distribution and the local behaviour of the generator mechanism.

The model is defined fixing the expected value of the sufficient statistics included in the probability distribution. Then, using the principle of maximum entropy, the most uncertain distribution is called *exponential random graph model* (or p^**-model*) because the joint distribution of the graph belong to the exponential family. Formally the graph G has distribution

$$P_\alpha(G) = \exp\{\alpha \cdot f(G) - \phi(\alpha)\}, \tag{3.22}$$

where $f(G)$ is a vector of sufficient statistics which are function of the adjacency matrix of G, like counting of various subgraph, degree distribution and so on.

The maximum entropy principle and the properties of the exponential family

give good reasons to use this model. Moreover, some choices of $f(G)$ have an interpretation in the description of the local stochastic process that describes the conditional dependencies between different edges. The downsize is the estimation process. Evaluate $\phi(\alpha)$ is often impossible, so the estimation is usually carried on with *Markov Chain Monte Carlo*, however there are issues because the method is still computationally very expensive. In the next chapter the theory behind models will be developed more deeply, in the last chapter $f(G)$ are moments of the eigenvalue distribution, i.e. the distribution of the closed walks in the network. A complete summary of exponential random graph models is Robins et al. (2007).

Chapter 4

Exponential Random Graph models

Introduction

Some graphs are similar to each other in some sense. For example, two possible networks may have similar average number of connected components, edges, triangles, cluster of vertices, eigenvalues or degree sequence. These characteristics are related to the *joint distribution* of the edges, which assign probabilities to all possible graphs. However, despite these quantities can be measured in real networks, the identification of the underline model is often really difficult, if not impossible. Moreover, some models that mimic a completely different generator mechanism can fit really well the data, because they lead to similar values for the descriptive statistics of the network. In chapter 2, the maximum entropy principle has been used to choose particular distributions for the data, which correspond to choose an exponential family. For this model, there is an explicit link between the distribution and a linear combination of sufficient statistics, which in the network case are functions of the adjacency matrix. The model is called *Exponential Random Graph Model* and the sufficient statistics are counts of particular subgraphs in the network.

Often in real networks the joint distribution depends on local processes, i.e. dependencies, that lead to the observed network. When this happens the edges ij and kh are dependent if and only if they are sufficiently close to each other. This is called *local level dependency*, if specified, lead to an explicit joint probability distribution of the network that belong to the exponential family, by the *Hammersley-Clifford theorem*. When the processes are "really local", meaning that only the neighbours of the vertices affect the probability of a tie between them, the model is called *Markov Random Graph Model*, because of the *Markov dependency* between actors.

4.1 Joint and Conditional Distributions

A *network model* is defined assigning probabilities in a reasonable way in the space of simple graphs. In this way a network is generated by a *stochastic process* that assigns the ties, the *observed network* is a realization of this process. Sometimes the stochastic process is influenced only by *local rules* between the vertices, which in this case are often called *actors*. In Robins et al. (2007), the network is defined as a "self organizing system of relational ties", the process models the mechanism of self organizing ties, and the observed network is assumed to be a realization of the process at one particular time. Of course, the specific realization is representative of the process if the last one is in some sort of thermodinamical equilibrium. When this is the case, the process generates networks with distribution close to the one with maximum entropy, with constraints the expected values of the quantities that govern the process.

Since the network model is governed by the local behaviour of the vertices, a big part of the information should be in small subgraphs, as they describe patterns of relations between close vertices. The subgraph counts are functions in the space $\mathcal{G}_n = \{0,1\}^{\binom{n}{2}}$, the set of simple graphs with n vertices, or equivalently, in the space of symmetric adjacency matrices. A k-dimensional sufficient statistic is $f : \{0,1\}^{\binom{n}{2}} \to \mathbb{N}^k$, and $f(G)$ is a vector of subgraph counts. A particular value of $f(G)$ is called *configuration*. The probability distribution in $\mathcal{G}_n$ defined by the sufficient statistics is a *Gibbsian ensemble*, so all networks with the same configuration are equally probable.

In this section G denote both the random and observed network, the random variables that correspond to the ties are $X = (X_{12}, X_{13}, ..., X_{n-1,n})$, where X_{ij} is the binary random variable associated with the tie between i and j. The first assumption is that $P(G) > 0$ for every $G \in \mathcal{G}$, which means that every edge appears in the network with positive probability, this is

$$P(G) > 0 \;\; \forall \;\; G \in \mathcal{G}_n \quad \Longleftrightarrow \quad \mathbb{P}(X_{ij} = 1) \in (0,1) \;\; \forall \;\; i \neq j. \tag{4.1}$$

Using the approach of chapter 2, the sufficient statistics $f(G) = (f_1, ..., f_m)$ are functions of the adjacency matrix of G, and they represent the information that we have on the network G. The expected values of the sufficient statistics are the constraints for the maximization of entropy, which for random graph can be written as

$$\begin{aligned} &\max_P \; -\sum_{G \in \mathcal{G}} \log(P(G))P(G) \\ &\text{s.t.} \sum_{G \in \mathcal{G}_n} f_i(G)P(G) = \mu_i, \quad \text{for } i = 1, ..., m, \\ &\quad P(G) > 0 \text{ for } G \in \mathcal{G}_n, \quad \sum_{G \in \mathcal{G}_n} P(G) = 1. \end{aligned} \tag{4.2}$$

The solution is

$$P_\alpha(G) = \exp(\alpha \cdot f(G) - 1 - \alpha_0) = e^{\alpha \cdot f(G) - F(\alpha)}, \tag{4.3}$$

where the parameter α_0 is included in the log partition function

$$F(\alpha) = 1 + \alpha_0 = \log\left(\sum_{G \in \mathcal{G}_n} \exp(\alpha \cdot f(G))\right). \tag{4.4}$$

The function F is the main obstacle of these model, because usually it can not be computed explicitly as the dimension of $\mathcal{G}_n$ is $2^{\binom{n}{2}}$.

In this model the observed network G affects the probability distribution only through the subgraph counts $f(G)$ which is a point in the vector space of the configurations. The Lagrange multipliers α are the canonical parameters of the family P_α, the parametric space is $\Theta \subseteq \mathbb{R}^m$.

Since P_α is a *Gibbsian ensemble*, the distribution over the *microstates* (networks) with same *macrostate* (configuration of sufficient statistics) is uniform. The distribution over the possible configurations belongs to the exponential family. This guarantees that networks with similar sufficient statistics are "close" to each other, but also that the sufficient statistics have all the information contained in the distribution, as different networks with same configuration are indistinguishable. P_α is therefore the most uncertain distribution, given the information that we have on the expected value of the sufficient statistics.

Maximize the entropy with respect to P is equivalent to minimize the relative entropy between P and a uniform prior distribution P_0 (in information sense). The entropy $H(P)$ is proportional to

$$D(P||P_0) = \sum_{G \in \mathcal{G}_n} P(G) \log\left(\binom{n}{2} P(G)\right), \tag{4.5}$$

which is the relative entropy of P with respect to the uniform distribution over the space of simple graphs

$$P_0(G) = \frac{1}{\binom{n}{2}}.$$

Note that the uniform distribution P_0 belong to the family P_α with $\alpha = 0$. In this case all possible networks have the same probability, so the pattern of ties can not contain information on the distribution of the network.

More interestingly, when $f_1(G) \propto |E(G)|$, equation 2.37 allows the decomposition

$$D(P_\alpha||P_0) = D(P_\alpha||P_{(\alpha_1,0)}) + D(P_{(\alpha_1,0)}||P_0), \tag{4.6}$$

because all three distributions belong to P_α. $P_{(\alpha_1,0)}$ is

$$P_{(\alpha_1,0)} = \exp\left(\alpha_1 |E(G)| - \binom{n}{2} \log(1 + e^{\alpha_1})\right). \tag{4.7}$$

If the expected number of edges is fixed to $\mu_1 = \mathbb{E}|E(G)|$, then

$$\mu_1 = \frac{d}{d\alpha_1}\binom{n}{2}\log(1+e^{\alpha_1}) = \binom{n}{2}\frac{1}{1+e^{\alpha_1}}, \tag{4.8}$$

and the probability of an edge between i and j in $V(G)$ is the constant

$$p = \frac{1}{1+e^{\alpha_1}} = \frac{\mu_1}{\binom{n}{2}}, \tag{4.9}$$

(Newman (2010), section 15.2), therefore $P_{\alpha_1,0}$ is the distribution of an Erdös-Rényi random graph.

The decomposition 4.6 can be interpreted in term of information. The second term $D(P_{\alpha_1,0}||P_0)$ is the information gained after knowing the expected number of edges, in this case the most uncertain distribution with this constraint is the Erdös-Rényi random graph, with probability p in equation 4.9. The first term $D(P_\alpha||P_{\alpha_1,0})$ is the difference in term of information of P_α from $P_{\alpha_1,0}$, which is the non-informative distribution in the case in which the expected number of edge is known. Given this sufficient statistic, the ERGM can be interpreted as the information gain from the Erdös-Rényi random graph, after that the expected values of the other sufficient statistics $f_2(G), ..., f_m(G)$ are known. Thus, the Erdös-Rényi random graph is the noninformative prior when only the expected average degree is known.

Suppose that G is has two connected components G_1 and G_2. Then there are no subgraphs with some vertices in G_1 and some in G_2. Therefore the counts f can be separated in f_1 and f_2 such that $f_1 + f_2 = f$, where f_1 and f_2 are the counts of the same subgraphs as f in G_1 and G_2 respectively. The distribution of G has density

$$P_\alpha \propto e^{\alpha\cdot f(G)} = e^{\alpha\cdot(f_1(G_1)+f_2(G_2))} = e^{\alpha\cdot f_1(G_1)}e^{\alpha\cdot f_2(G_2)}. \tag{4.10}$$

This shows the third axiom of Shore and Johnson (1980) introduced in section 2.2. In this case the network can be decomposed two independent subsets (two ties in different components are conditional independent, given the rest of the network) and the inference on α can be done separately.

Conditional distribution

Equation 4.3 is the joint probability distribution function of an exponential random graph model. Now is exposed the relation between 4.3 and the behaviour of the graph at the local scale. Denote with G^+_{ij} the network G when the edge ij is forced to be present, G^-_{ij} when ij is absent. Denote with G^c_{ij} the set $E(G)\setminus ij$, which is the observed network G without the edge between i and j, with regard ij is present or not.

The local behaviour affects the substructures in the generated graph, the probability of an edge can be modelled such that depends by a linear combination of substructures. Equivalently, to have an unbounded linear term, consider

the logarithm of the odds

$$\log(\text{odds}(X_{ij}=1)) = \log \frac{\mathbb{P}(X_{ij}=1)}{\mathbb{P}(X_{ij}=0)} = \log \frac{\mathbb{P}(X_{ij}=1)}{1-\mathbb{P}(X_{ij}=1)}. \tag{4.11}$$

Now, the log-linear model is

$$\log \frac{\mathbb{P}(X_{ij}=1)}{1-\mathbb{P}(X_{ij}=1)} = \alpha \cdot f(X)_{ij} \quad \Longleftrightarrow \quad \mathbb{P}(X_{ij}=1) = \frac{\exp(\alpha \cdot X_{ij})}{1+\exp(\alpha \cdot X_{ij})}, \tag{4.12}$$

where $f(X)_{ij}$ is a function of the adjacency matrix which has to be defined.

In summary, we know the joint probability distribution of the network and the marginal distribution of an edge. Since the ties are dependent, in Van Der Pol (2017) and Strauss and Ikeda (1990) is suggested to use the *conditional odds*

$$\log \frac{\mathbb{P}(X_{ij}=1|G_{ij}^c)}{\mathbb{P}(X_{ij}=0|G_{ij}^c)} = \alpha \cdot f(G_{ij}^c). \tag{4.13}$$

Then the function $f(G_{ij}^c)$ can be fixed considering the joint distribution 4.3. Since the sufficient statistics (quantities that contains all information on the distribution) are subgraphs, the function $f(X)_{ij}$ has to be related to sub-network counts. G_{ij}^c is fixed, so are the subgraphs in G_{ij}^c. If $X_{ij}=0$, the sub-networks are the same as in G_{ij}^c, so $f(G_{ij}^c) = f(G_{ij}^-)$. Then the model equivalent to 4.3 is

$$\log \frac{\mathbb{P}(X_{ij}=1|G_{ij}^c)}{1-\mathbb{P}(X_{ij}=1|G_{ij}^c)} = \alpha \cdot (f(G_{ij}^+) - f(G_{ij}^-)). \tag{4.14}$$

The parameter α regulates the average change of sufficient subgraphs when one edge is modified.

The last equation can be used to generate a *Markov Chain* of networks in which the difference between two successive graphs is at most one edge. This is the main tool in estimation, as we can choose an algorithm which doesn't need to compute the log partition function $F(\alpha)$. Moreover, the Markov Chain can be used to mimic the unknown process that generates the network, at least in an approximate way such that at each time, one tie can be modified. The assumption is that the network is observed in a time such that the generative process is in thermodynamic equilibrium, the Markov Chain is often a good indicator if this assumption can be reasonable. More details about the estimation process will be given later in this chapter.

4.2 Sufficient Statistics and Local Dependencies

The model 4.3 specifies a distribution over the space of simple random graphs. Often in real networks this distribution depends on local processes, i.e. dependencies, that lead to the observed network. When this happens the edges X_{ij} and X_{kh} are dependent if and only if they are sufficiently close to each other.

This is called *local level dependency*, if specified, leads to an explicit joint probability distribution of the network, that belongs to the exponential family, by the *Hammersley-Clifford theorem*.

If the assumption 4.1 is fulfilled, every network model can be written as

$$P(G) \propto \exp \Lambda(G), \tag{4.15}$$

with $\Lambda(G) = \log P(G) - \log P(G^*)$. For a fixed graph $G^* \in \mathcal{G}_n$. On the other way around, a network model can be defined with density as in 4.15 for every $\Lambda : \{0,1\}^{\binom{n}{2}} \to \mathbb{R}$. The Hamiltonian $\Lambda(G)$ is a log-ratio and the problem is binary, so the reference configuration can be chosen to be the empty graph: $G^* = 0_n$. Every network distribution is so a log ratio between the observed network G and the empty graph.

In Frank and Strauss (1986)) is shown that the graph Hamiltonian $\Lambda(G)$ can be written as

$$\Lambda(G) = \log P(G) - \log P(0_n) = \alpha \cdot (f(G) - f(0_n)), \tag{4.16}$$

and f can be chosen such that $f(0_n) = 0$. This choice allows to use only subgraph counts as sufficient statistics of every network, because the difference $f(G) - f(0_n) = f(G)$ depends only on the edges of G, the location of non-edges is not important. Therefore $f(G^*) = f(0_n) = 0$ is the reference point in the vector space of log densities as is shown in section 2.4 and the parameters α are the coordinates of the vector $f(G)$ in this space. In order to choose which subgraphs include in the Hamiltonian, we need to specify which sets of edges are conditional dependent.

The *(local) dependence assumption* is a rule in $E(G) \times E(G)$ that assigns conditional dependencies between ties, i.e. fix which couples of edges can affect each other, given the rest of the network. For example in the *Markov dependence* exposed below, two ties are conditionally dependent if they are incident (they share a vertex). Following Frank and Strauss (1986), let's define the *dependence graph* $D(G)$ as the network with $\binom{n}{2}$ vertices, which correspond to the edge set of G and they are denoted as $V(D(G)) = \{12, 13, ..., (n-1)n\}$, the edge set of the dependence graph is

$$(ij)(kl) \in E(D(G)) \;\text{ if } \{ij, kl\} \subset E(G), \tag{4.17}$$

with $1 \leq ... \leq i < j \leq ... \leq n$ and $1 \leq ... \leq k < l \leq ... \leq n$. Note that the dependence graph usually have self-loops, but these can be excluded as the self-loop implies the obvious fact that an edge is conditionally dependent to itself. In figure 4.1 there is a network G and its associated $D(G)$ with Markov dependence assumption.

After the dependency rule is chosen, there is an explicit relation between the dependence graph and the sufficient count statistics of the ERGM. In particular, the dependence assumption gives an explicit formula for $\Lambda(G)$, the log-ratio between the observed network and the empty graph. The *Hammersley-Clifford*

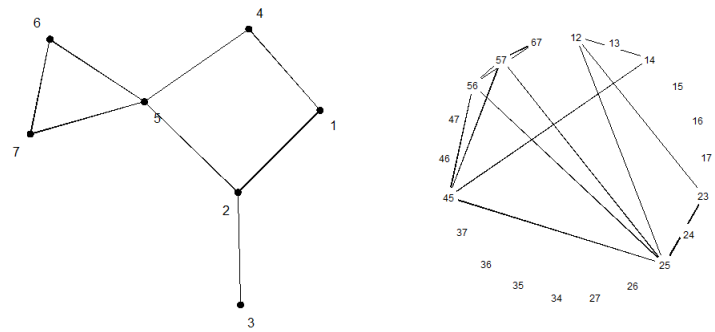

Figure 4.1: Network G and correspondent $D(G)$ with Markov dependency: two edges are conditionally dependent if they share a vertex.

theorem (Besag (1974)) implies that $\Lambda(G)$ is a linear combination of the subgraphs of G associated with the cliques of $D(G)$.

Theorem 1. (*Hammersley-Clifford*) Any simple graph G, with associated dependence graph $D(G)$, can be written as

$$P(G) \propto \exp \sum_{c \in \widetilde{\mathscr{C}}(D(G))} \alpha_c, \tag{4.18}$$

where $\widetilde{\mathscr{C}}(D(G))$ is the set of maximal cliques of the dependence graph $D(G)$.

There is an explicit relation between some subgraphs of G and the maximal cliques of $D(G)$. For example, in figure 4.1 $D(G)$ is computed using Markov dependency. The clique (12)(23)(25) correspond to the 3-star centred in 2, (25)(45)(56)(57) correspond to the 4-star centred in 5, which contains itself the 3-stars correspondent to the cliques (25)(45)(56), (25)(45)(57), (25)(56)(57) and (45)(56)(57). The triangle $5, 6, 7$ correspond to the clique (56)(57)(67).

Since the vertices are unlabelled, subgraphs of G (cliques of $D(G)$) of the same type are equivalent in the distribution. This is called *homogeneity condition*: as the model is invariant under permutation of the labelling, the parameters α_c in equation 4.18 are not affected by which actors form the subgraph. This type of graph is called *homogeneous network*. For example, in figure 4.1 there are four 3-stars, (three of them centred in 5 and one in 2). The parameters α_c are equal for all the cliques c that correspond to a 4-star. The Hammersley-Clifford theorem combined with the homogeneity condition, defines every possible network model invariant under permutation of the vertices.

The family distribution of homogeneous networks is so

$$P_\alpha(G) = \exp\left(\sum_{c\in\mathscr{C}(D(G))} \alpha_c|\mathscr{S}_c(G)| - \psi(\alpha)\right). \tag{4.19}$$

where $\mathscr{C}(D(G))$ is the set of cliques such that each of them correspond to a different subgraph $\mathscr{S}_c$. Note that because of 4.15, every network distribution invariant under relabelling can be written in this way. Therefore, with the subscript α is emphasized that P_α is a family of distribution over the space of homogeneous simple networks, which is actually an exponential family in which the sufficient statistics are the subgraphs counts associated with the cliques in the dependence graph. Every fixed value α^* specifies a distribution in this family, i.e. a random graph. The parameters α lie in the canonical parametric space $A \subseteq R^{|\mathscr{C}(D(G))|}$.

It's reasonable to assume that a relevant part of the information can be explained by "small" subgraphs, like triangles, stars and so on, in this way only a finite amount of parameters α_c are non-zero. The properties of the exponential family imply that all the information contained in the distribution, is contained in the sufficient statistics $\mathscr{S}_c(G)$, when $\alpha_c \neq 0$.

ERGM for Nonhomogeneous and Directed Graphs

While homogeneity of a network is easy to define, using invariance under relabeling, non-homogeneity can take very different forms. In principle every network model can be represented using $P(G)$ in equation 4.18, but this specification involves way too many parameters, as all possible conditional dependencies between actors is included. If there are information on the specific vertices, this form of non-homogeneity can be modelled including this information as covariates in the vertex set, this approach is not developed in this thesis, except for the simplest case in which there is only one covariate that denotes the community in which the vertex belongs.

We can then reduce the number of parameters in $P(G)$ in the following way. We need to introduce counts of equivalent subgraphs, but in this case not only their form is important, but also the labels of the vertices that compose them. The Hammersley-Clifford implies that the model takes the form

$$P_{\alpha;k} = \exp\left(\sum_{c\in\mathscr{C}(D(G))}\sum_{j\in J_c} \alpha_{c,j}|\mathscr{S}_{c,j}(G)| - \psi_k(\alpha)\right), \tag{4.20}$$

where $|\mathscr{S}_{c,j}(G)|$ is the count of the subgraph $\mathscr{S}_c$ with labels $j \in J_c$ defined as

$$J_c = \left\{j_1, ..., j_{|V(\mathscr{S}_c)|} \text{ s.t. } j_i \in \{1, ..., k\}\right\}, \tag{4.21}$$

$|V(\mathscr{S}_c)|$ is the size of the subgraph $\mathscr{S}_c$. The number of parameters is still too large but, with some dependence assumptions, the model can be approximated

reasonably well.

Exponential random graph models can be used also for directed networks. In this case the edges have a direction, therefore (if self loops are not allowed) there are in total $n(n-1)$ possible edges. The *directed Erdös-Rényi* random graph is the usual extension of the undirected version, in which the only parameter fix the probability p of an edge between two vertices. In this model, the probability of a *dyad* (there is an edge from i to j and one from j to i) is p^2. Also the binomial random graph can be extended with the same approach, the number of parameters is $n(n-1)$, one for each potential edge.

The first nontrivial ERGM for directed graph is the *dyad independent random graph*, in which the edge from i to j and the one from j to i are conditionally dependent, while all other pairs are conditionally independent. In this model there are two parameters, one for the edge density, and the other to model the dependence between edges that connect the same vertices, which it's usually positive. In fact if $i \to j \in E(G)$, then it's likely that $j \to i$ is also in $E(G)$ (if i considers j his friend, it's likely that also j considers i as friend). Also for directed networks more complicated dependence assumptions can be used, in the next section is introduced the random graph with Markov dependency, both for undirected and directed networks.

4.3 Markov Random Graphs

The most local dependence assumption is the *Markov dependency*, the relation between i and j is influenced only by their neighborhoods and the intersection between them. At edge level, the subgraphs associated with neighborhoods are the k-stars, $k = 1, ..., n-1$ and the one associated with intersection between neighborhoods of two vertices is the triangle. The *Markov Random Graphs* has distribution

$$P_{\tau,\theta}(G) \propto \exp\left(\tau|C_3(G)| + \sum_{i=1}^{n-1} \theta_i |S_i(G)|\right), \tag{4.22}$$

where $|C_3(G)|$ is the number of triangles and $|S_i(G)|$ is the number of i-stars. Of course the distribution is both in form 4.19 and 4.3.

Using the reparametrization

$$|S_i| = \sum_{h \geq i} \binom{h}{i} \tilde{d}_h \quad \text{and} \quad \mu_h = \sum_{i \leq h} \binom{h}{i} \theta_i, \tag{4.23}$$

where $\tilde{d}_h$ is the proportion of vertices with degree h, the distribution 4.22 can be written as

$$P_{\tau,\mu}(G) \propto \exp\left(\tau|C_3| + \sum_{h=1}^{n-1} \mu_h \tilde{d}_h\right). \tag{4.24}$$

In this form the parameter μ_h control the bias for or against vertices of degree h (Frank and Strauss (1986)). This model is an ERGM, and is the distribution

with maximum entropy with given *expected degree distribution*. In many real cases, especially for network which are large, the variance of the average degree is unbounded, like in the power law (equation 3.21) when $2 < \alpha \leq 3$. When this happens, the sufficient statistics can have very large values and the model is not useful (more generally, the maximization of entropy is not constrained by an inequality if its variability is very large or unbounded).

The correspondent conditional distributions can be written explicitly and is

$$\log \frac{\mathbb{P}(X_{ij} = 1 | G_{ij}^c)}{1 - \mathbb{P}(X_{ij} = 1 | G_{ij}^c)} = \tau \Delta(C_3) + \mu_{(i)} + \mu_{(j)}, \tag{4.25}$$

where $\Delta(C_3) = |C_3(G_{ij}^+)| - |C_3(G_{ij}^-)|$, $\mu_{(i)}$ is the parameter associated with $\tilde{d}_{(i)}^+$, the proportion of vertices with degree equal to i when $ij \in E(G)$. If i and j have the same degree, then $\mu_{(i)} = \mu_{(j)}$.

For the next discussion, without loss of generality it can be assumed that $\tau = 0$, so the triangles are not relevant in the model. Therefore the Hamiltonian of the Markov random graph depends only on the degree distribution $\tilde{d}$. Let $\varphi : \mathcal{D} \to \mathbb{R}$ a function defined in the space of degree distributions $\mathcal{D}$, which is a vector space because every degree distribution is a n dimensional vector with entries in $\{0, ..., n-1\}$. If φ is bounded, then

$$\varphi(\tilde{d}) = \sum_{h=1}^{n-1} \mu_h^* \tilde{d}_h = \sum_{i=1}^{n-1} \theta_i^* |S_i(G)| \tag{4.26}$$

(Snijders et al. (2006), section 3), The parameter vectors μ^* and θ^* depend on φ and are related by 4.23.

We can consider approximate models using the geometrical properties of the exponential family introduced in section 2.4. If the first parameters $\mu_1, ..., \mu_t$ are fixed to 0, the model

$$Q_{\mu,t} \propto \exp\left(\sum_{h=t+1}^{n-1} \mu_h \tilde{d}_h \right) \tag{4.27}$$

is a $n-1-t$ dimensional approximating family of $P_{0,\mu}$ in which only vertices with degree higher than $t+1$ are important in the model. The quality of the approximation depends on how important are the vertices with low degree.

Markov random graphs can be extended to nonhomogeneous networks with community structure. In this case the Hammersley-Clifford theorem implies that we need to counts the subgraphs distinguishing the labels of the vertices. Suppose that there are k communities, the degree of the i-th vertex in this case is a vector $di = (d_{i(1)}, ..., d_{i(k)})$, where $d_{i(j)}$ is the number of neighbours of the i-th vertex in the j-th community. Denote with $\mathcal{D}_k$ the vector space of degree distributions with vertices belonging to k communities. In principle every function $\varphi_k : \mathcal{D}_k \to \mathbb{R}$ can be written using an Hamiltonian like for homogeneous network, however the number of parameters grows very quickly.

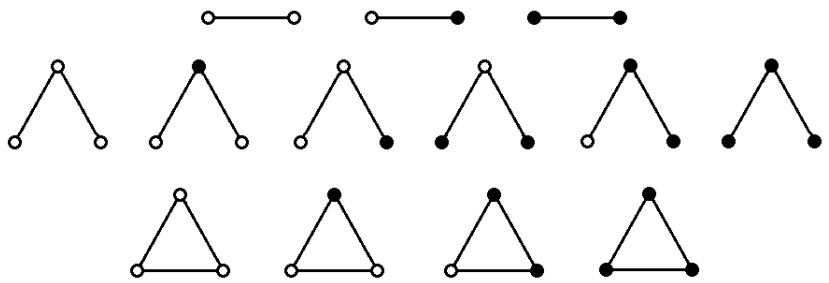

Figure 4.2: edges, 2-stars and triangles with two communities of vertices.

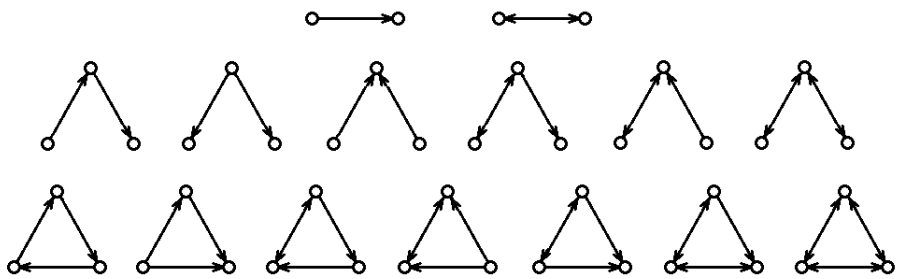

Figure 4.3: Directed subgraphs with less than 3 vertices.

In this set up is so necessary to consider approximate models, also when there are few communities. For example, with two communities the smallest relevant subgraphs are in figure 4.2, in this case there are 13 relevant parameters.

Markov dependency can be used also for directed networks but the number of parameters tend to grow quickly. In fact for every undirected subgraph, there are many combinations of directions of the edges. Likewise Markov graphs with community structure, all useful models are approximate, as only the small subgraphs are included in the Hamiltonian. In figure 4.3 there are all possible subgraphs with less than 3 vertices, the double arrow represents the two edges $i \to j$ and $j \to i$. The model with subgraphs in the figure has 15 parameters.

Markov random graphs in canonical form, like 4.22 or 4.24 are not useful because the Hamiltonian is too flexible. This is an usual problem of discrete exponential families, and it's the main topic of the next section.

4.4 Theoretical and Computational Problems

Most of the literature on exponential random graph models heavily emphasizes their computational problems. This is not unexpected, as most models in canonical form, such 4.3, 4.19 or 4.22, simply doesn't work. However, many

computational issues, such the problems in convergence of the Markov chain for example, are caused by some theoretical characteristics of discrete exponential families. The theoretical problems of these models are *instability*, *sensitivity* and *near-degeneracy*. The first one, if present, implies the two others. It seems that these characteristics are not pathological, meaning that they don't invalid the whole approach behind ERGM. In fact, when these characteristics are taken into account in the development of the model, most issues are solved.

Another difficulties in working with exponential random graph models is the consistency of the estimates. Sometimes the network is relatively small so the limit behaviour is not an issue. However, in any case in which the observed network is sampled from a (much) larger graph, consistency is critical. The thermodynamic interpretation of the exponential random graph model does not help, as the generative process does not model how the graph "grows". In particular, consistency is important when we are interested to make the results of the analysis independent on the size of the network. Unfortunately, exponential random graph models seem not to be consistent in the graph space.

Instability and Related Properties

In short, there are constraints in the configuration space given by the fact that the system is a network, rather than a variable in which all possible configurations in $\mathbb{N}^k$ are allowed. Therefore, the canonical parametric space A is always too large, and the mean value $M = \mu(A)$ is bigger than the space of allowed configurations (which is not even convex). However for Markov random graphs, one of the most general and flexible ERGM, has been found a lower dimensional curved families in which most of these issues are solved and it seems that these parametrizations perform well in practise (Snijders et al. (2006)). Despite so, a general procedure to specify an exponential random graph models without worry of theoretical/computational issues is not yet known, so they need to be taken into account when a model is used in practise.

Following Schweinberger (2011), the first notion introduced is instability. Let's rewrite the exponential random graph model in a way such that the size of the model is taken into account:

$$P_\alpha(G_n) = e^{\theta_N(\eta)\cdot f_N(G_n) - F_N(\eta)}, \tag{4.28}$$

where $N = n(n-1)/2$ are the *degrees of freedom* and G_n is an n dimensional graph. A canonical family can be written in a way that θ does not depend on the dimensionality, however the form 4.28 gives more insight to the problem.

A discrete exponential family is *stable* [*unstable*] if exist $C > 0$ and $N_C > 0$ such that

$$\max_{G_n \in \mathcal{G}_n} \theta_N(\eta) \cdot f_N(G_n) \;\leq\; [>] \;\; CN \quad \text{for all } N > N_C. \tag{4.29}$$

This is a constraint on the asymptotic maximum size of the Hamiltonian, which can grow with a rate of at most $O(N) = O(n^2)$. The dependence of the size in

$\theta_N(\eta)$ has been emphasized because the instability of an Hamiltonian depends on the instabilities of the sufficient statistics that compose them. However standardize $f_{N,j}$ dividing it by $\max_{G_n} f_{N,j}(G_n)$ is useless because this transformation is an isomorphisms, so the problem has just been moved in the parametric space, instead of the configuration space. As stated before, if the family is unstable, then is also too sensitive and near degenerate. Therefore unstability is a necessary condition for obtaining a useful model.

Consider the n dimensional graphs $G^+_{n,ij}$ and $G^-_{n,ij}$ equal except in the edge ij, which can be seen an update of the Markov Chain Monte Carlo. The logit in equation 4.14 is equal to $\theta_N(\eta) \cdot (f_N(G^+_{ij}) - f(G^-_{ij}))$. If the model is unstable, then exist no $C > 0$ and $N_C > 0$ such that

$$|\theta_N(\eta) \cdot (f_N(G^+_{ij}) - f(G^-_{ij}))| \leq C \tag{4.30}$$

for all possible graphs that differs of one edge. This condition is called *sensitivity*, if n is large some log-odds are unbounded.

Lastly, consider the set of *modes*

$$\mathcal{M}_{\epsilon,N} = \left\{ G_n : P(G_n) > (1-\epsilon) \max_{G_n \in \mathcal{G}_n} \theta_N(\eta) \cdot f_N(G_n) \right\}, \tag{4.31}$$

where $N = n(n-1)/2$ are the degrees of freedom. The distribution is *near-degenerate* if

$$\mathbb{P}_\theta(G_n \in \mathcal{M}_{\epsilon,N}) \to 1. \tag{4.32}$$

When this happens most of the probability mass is concentrated on a few configurations, usually very different than each other. If the model is unstable, then is near-degenerate (Schweinberger (2011), theorem 2). In section 2.3, has been shown that if the exponential family is minimal, the mean value parameter lies in $M = \text{int}(K_\nu)$, the interior of the convex hull of the support. In Handcock et al. (2003) is shown that when the model is near degenerate, the mean value parameter tend to the boundary of M. Moreover, in the same paper is shown that degenerate models diverge in relative entropy from the non-degenerate models.

A solution that seems to work well in practise is to consider *curved exponential families*, so distributions in which the parametric space is a curved low-dimensional subspace of A. This can force the expected values of the sufficient statistics to be "far away" from the boundaries of the support. Consider the Markov random graph defined in equation 4.24, for simplicity assume $\tau = 0$ so that the Hamiltonian is a linear combination of the degree distribution. We need to impose constraints so that $\theta_N(\eta) \cdot f_N(G_n)$ does not grow too quickly with n, for Markov random graphs this Hamiltonian is influenced mostly by the vertices with high degree. A solution is to use a *geometrically decreasing degree distribution*

$$u^{(d)}_\alpha(G) = \sum_{h=1}^{n-1} e^{-\alpha h} \tilde{d}_h(G), \tag{4.33}$$

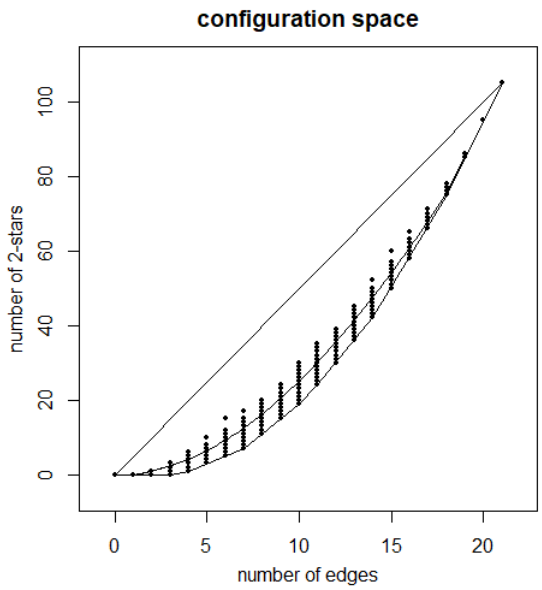

Figure 4.4: Possible configurations (dots) of edges and 2 stars for graphs with 7 nodes. The space is not convex, so the mean value parameter is interior of the convex hull of the dots. The curve that passes through the dots is the space of sufficient statistics in a curved Hamiltonian, forcing the model to have expected sufficient statistics in the configuration space.

which is equivalent to

$$u_\lambda^{(s)}(G) = \sum_{i=2}^{n-1} (-1)^i \frac{|S_i(G)|}{\lambda^{i-2}}. \tag{4.34}$$

The model can be written in canonical form as in equation 4.22 and 4.24 using

$$\mu_h(\alpha) = e^{-\alpha h} \quad \text{and} \quad \theta_i(\lambda) = \frac{(-1)^i}{\lambda^{i-2}}. \tag{4.35}$$

Therefore for scaling exponentially the degrees we need to counterbalance the effect of the k-stars. In fact $|S_i(G)|$ and $|S_{i+1}(G)|$ are heavily correlated, so $\theta_i(\lambda)$ and $\theta_{i+1}(\lambda)$ have a different sign. The model is stable if $\alpha > \log(2)$ and therefore $\lambda = e^\alpha/(e^\alpha - 1) > 2$.

Curved exponential random graph models are introduced in Snijders et al. (2006), in the paper there are also other examples of curved Hamiltonians that can be used in practice. In figure 4.4 is shown why this approach is useful. Forcing the Hamiltonian to stay in the support of the family we can avoid degeneracy, when the parametrized curve is far away from the boundaries of the support. The statistics 4.33 and 4.34 can be computed quickly after the update of an edge, this is useful because the estimation uses a Markov Chain Monte Carlo.

Using the conditional distribution 4.14, if at time t, ij is updated to 1 in the

Metropolis-Hastings algorithm, the update is accepted with probability

$$\mathbb{P}_\alpha(X_{ij}^{(t+1)} = 1 | G_{ij}^c) = \min\left\{1, e^{\alpha\cdot(f(G_{ij}^+)-f(G_{ij}^-))}\right\}. \tag{4.36}$$

If ij is updated to 0, the formula is the same with $f(G_{ij}^-) - f(G_{ij}^+)$. The advantage of this algorithm is that we don't need to compute the partition function $F(\alpha)$. This algorithm allows us to generate samples from P_α, so the chain can be interpreted as an approximation of the generative thermodynamic process. If we observe a network, we can make inference using the *estimated generative process*, which is obtained from the chain with stationary distribution $P_{\hat{\alpha}}$.

The algorithm is used also for the estimation which, by the properties of the exponential family, is a convex problem. In the estimation, the chain 4.36 is used with parameter $\hat{\alpha}^{(m)}$, for $m = 1, ..., M$. Then the m-th sample is used to update $\hat{\alpha}^{(m)}$ to $\hat{\alpha}^{(m+1)}$ using a convex optimization algorithm, like *Newton-Raphson* for example. An optimization procedure is described in more detail in Snijders (2002). This method of inference is the usual procedure with dependent data, see Geyer and Thompson (1992). Examples of these algorithms in R are in Van Der Pol (2017).

Consistency of ERGMs

The main issue of exponential random graph models is consistency. In fact, the generative stochastic process has been specified fixing the number of vertices. However, when $n \to \infty$ we need to specify a model that describes how the graph grows and the convergence of sequences in a graph space is complicated. This theory is called *graph limit*, the main reference is Lovász (2012). However, the theory is applicable only for dense networks (average degree $O(n)$). For dense exponential random graph models consistency has been studied in Chatterjee et al. (2013).

Let $G_1, G_2, ...$ a sequence of graphs of increasing dimension. If the elements of the sequence are dense, then G_n converges to a continuous function $h : [0,1]^2 \to [0,1]$ called *graphon*. This function is symmetric in its two arguments, and it can be chosen to be nondecreasing, in the sense that $h(x,y) \leq h(x, y+\epsilon)$, if $\epsilon > 0$ and $y + \epsilon < x$. Every function that satisfies these properties is obtained by a sequence of graphs for a specific probability distribution. Therefore, the *graphon space* $\mathcal{H}$ is the appropriate limit object for sequences of dense random graphs. For exponential random graph models, the distribution of G_n depends on the sufficient statistics

$$f_i(G_n) = \frac{t(S_i, G_n)}{n^{|V(S_i)|}}, \tag{4.37}$$

where $S_1, ..., S_m$ are subgraphs and $t(H, G)$ counts the occurrences of the subgraph H in the network G.

The counts are rescaled by $n^{|V(S_i)|}$ in order to have a nontrivial limit in the

dense regime. This limit is

$$\tilde{f}_i(h) = \lim_{n\to\infty} f_i(G_n). \tag{4.38}$$

For n finite, the distribution of G_n is obtained by the graphon h, with a partition of $[0,1]^2$ in n^2 blocks. The limit object of the exponential random graph model is

$$p_n(h) = e^{n^2\theta\cdot\tilde{f}(h)-\phi_n(\theta)}, \tag{4.39}$$

where

$$\phi_n(\theta) = \frac{1}{n^2}\sum_{\mathcal{G}_n} e^{n^2\theta\cdot\tilde{f}(h)}. \tag{4.40}$$

If $f_1(G_n) = n^{-2}|E(G_n)|$ and $\theta_2, ..., \theta_m$ are nonnegative, then

$$\lim_{n\to\infty}\phi_n(\theta) = \phi_\infty = \sup_{0\le u\le 1}\left(\sum_{i=1}^{m}\theta_k u^{|E(S_i)|} - I(u)\right), \tag{4.41}$$

where

$$I(u) = \frac{1}{2}u\log u + \frac{1}{2}u\log(1-u), \tag{4.42}$$

and ϕ_∞ is constant. Moreover, if the maximization 4.41 has one solution u^*, then G_n is indistinguishable from an Erdös-Rényi graph Chatterjee et al. (2013).

These results seem to preclude strong forms of consistency, like in almost sure sense. A particular form is *consistency under sampling*, useful when the observed network is sampled form a larger graphs. The graphon is interpreted as an infinite dimensional dense networks from which the (dense) observed one is sampled. Stochastic block models, the main non-parametric statistical approach for network data, perform very well in this context, as the graphon can be estimated consistently, Airoldi et al. (2013). In Shalizi and Rinaldo (2013) is shown that consistency under sampling is a problem for some types of exponential families for dependent random variables, which include ERGMs.

In information context, the interest is on the probability distribution of the system. In particular, we may only been interested on the weak consistency of $\hat{\alpha}$ as estimator of α, for example if we have an hypothesis on the properties of the generative process in a macro scale. In this setting, exponential families are usually well behaved. However, for weak consistency we have at least to specify how the expected value and variance of the sufficient statistics grow with n in order to have a non-trivial limit for $\hat{\alpha}$.

In the next chapter the assumption of Markov dependency (and associated degree distribution as sufficient statistic) is dropped. The exponential random graph model introduced in the next chapter has sufficient statistics related to the eigenvalue distribution.

Chapter 5

Exponential Random Graph Models with Spectral Statistics

Introduction

Exponential random graph models have been introduced starting from the local dependence assumption, which specifies the relevant subgraphs in the network distribution, that belongs to the exponential family. Markov dependency is the most used, because it is particularly well behaved mathematically, and it can be used both in homogeneous networks, but also in inhomogeneous graphs with community structure. More general dependencies are way less tractable, however for homogeneous networks, spectral moments of the adjacency matrix can be used as sufficient statistics. As introduced in chapter 3, these quantities are associated with number of closed walks in the network. In this section it's used an exponential random graph model as approximating family of a general homogeneous random graph model, with family of distributions specified by a general dependence rule. In the approximating model the sufficient statistics are number of closed walks. If the dimension of the graph is fixed and we use enough sufficient statistics, i.e. we consider counts of closed walks up to a length big enough, the approximating model is equivalent to the original, so there is no information loss.

5.1 Higher Order Local Dependencies

Markov random graphs use as configurations triangles and k-stars, however many other subgraph counts may be important as sufficient statistics. Looking the tie placement near the edge ij, Markov dependency assumptions is that only actors in the neighborhood of i and j affect the probability that the tie ij is in

the network. This may be not enough, as the presence of ij can be influenced by the tie placement in a bigger neighborhood around ij.

A way to define this type of *extended neighbour* of the edge ij is through the use of closed walks that pass through ij. Let $N_k(ij)$ the set of edges $hl \in E(G)$ such that exist a closed walk of k step that cross both ij and hl (note that ij always belong to $N_k(ij)$). If the total number of k-step closed walks that cross ij is $W_k(ij)$, then the importance of every $hl \in N_k(ij)$ in determinate the presence of ij is given by the proportion $W_k(ij;hl)/W_k(ij)$. Therefore the weighted influence of hl in ij can be defined as

$$\sum_{k=2}^{m} \alpha_k \frac{W_k(ij;hl)}{W_k(ij)}. \tag{5.1}$$

Analogous formulas can be derived at vertex level, considering the number of k-step closed walks $W_k(i)$ and $W_k(i;j)$, that cross the vertex i and the vertices i and j respectively.

A dependence assumption specifies the vector of parameters α in the family 4.19. In particular a dependence rule allows a finite amount of $\alpha_c \in \alpha$ to be different than 0. However, in a dependence rule in which some subgraphs are specified as relevant, there may be cliques in the dependence graphs that are associated with subgraphs specified as non-relevant, but the Hammersley-Clifford theorem force their counts to be important in P_α. For example, in figure 5.1 the dependence rule has been defined such that only the number of edges and C_4 are relevant, but because of the Hammersley-Clifford theorem, all other subgraphs are relevant, including the one that represent the whole structure. The graph in the figure is a 3-triangle and often appears in real networks. The only subgraphs that does not add new relevant structures are k-stars and triangles, the ones considered for the Markov dependency.

The Hammersley-Clifford theorem and the homogeneity assumption define the original family of distribution

$$P_\alpha(G) = e^{\alpha \cdot \mathscr{S}(G) - \psi(\alpha)}, \tag{5.2}$$

where $\psi(\alpha)$ is the log-partition function and $\alpha \cdot \mathscr{S}(G)$ is the *graph Hamiltonian*

$$\sum_{c \in \mathscr{C}(D(G))} \alpha_c |\mathscr{S}_c(G)|, \tag{5.3}$$

$\mathscr{C}$ is the set of different maximal cliques, $|\mathscr{S}_c(G)|$ is the count of the subgraph associated with the clique c. Therefore, the exponential family 5.2 is characterized by the vector of sufficient statistics

$$\mathscr{S} = \left(|\mathscr{S}_1|, ..., |\mathscr{S}_{|\mathscr{C}|}|\right) \in \mathbb{N}^{|\mathscr{C}|}, \tag{5.4}$$

which are subgraph counts. When $n = V(G)$ is fixed, $|\mathscr{C}|$ is finite and $|\mathscr{S}_c|$ are bounded for all c.

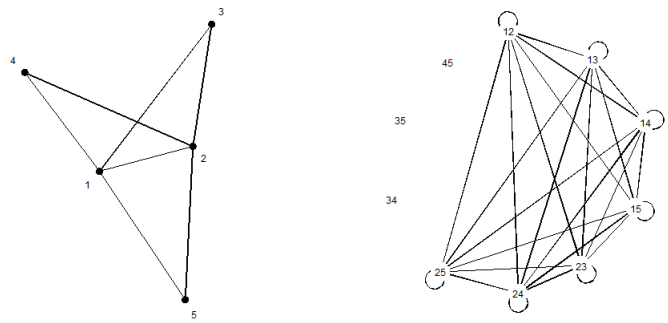

Figure 5.1: 3-triangle and associated dependence graph, with this dependence rule all subgraphs are relevant in the distribution of G.

The family P_α is invariant under bijective linear transformations of the sufficient statistics. Let $\varphi_{\mathbf{B},b} : \mathbb{R}^{|\mathscr{C}|} \to \mathbb{R}^{|\mathscr{C}|}$ defined as

$$\varphi_{\mathbf{B},b}(\mathscr{S}) = \mathbf{B}\mathscr{S} + b, \tag{5.5}$$

where b and $\mathbf{B}$ are a fixed vector and nonsingular matrix respectively. The Hamiltonian remains the same if α is modified as

$$\varphi^{-1}_{\mathbf{B},b}(\alpha) = \mathbf{B}^{-1}\alpha - a, \tag{5.6}$$

note that a does not need to be specified as the $\alpha \cdot b$ can be included in the partition function. Therefore, the model P_α is equivalent to all other families obtained with isomorphisms in the space of sufficient statistics.

A lower dimensional family can be used as "low rank" *approximating distribution* of the real family P_α. The transformation is $W = \varphi_{\mathbf{W}}(\mathscr{S}) = \mathbf{W}\mathscr{S}$, where $\varphi_{\mathbf{W}} : \mathbb{R}^{|\mathscr{C}|} \to \mathbb{R}^{m-1}$. W is the vector of counts of closed walks, with elements

$$W_k(G) = \mathrm{tr}(\mathbf{A}(G)^k) = \sum_{i=1}^{n} \lambda_i^k, \tag{5.7}$$

where $\mathbf{A}(G)$ is the adjacency matrix of G, $k = 2, ..., m$. Therefore W can be used as sufficient statistics for a minimal exponential random graph model of order $m-1$. The approximating family is

$$Q_\theta(G) = e^{\theta \cdot W(G) - \phi(\theta)}, \tag{5.8}$$

where $\theta \cdot W = \sum_{k=2}^{m} \theta_k W_k$. Later will be shown that the Hamiltonian $\theta \cdot W$ and the log partition function $\phi(\theta)$ are explicit functions of $\alpha \cdot \mathscr{S}$ and $\psi(\alpha)$.

There are some technical assumptions for identificability, they are not necessary because every nonminimal family can be reduced to a minimal one. Moreover, equation 2.26 implies that there are no problems with the inference, after reduction to a minimal sufficient statistic. Thus it can be assumed that Q_θ is minimal, so $\Theta \subseteq \mathbf{R}^{m-1}$ with full dimension (it contains an $m-1$ dimensional open sets). If we assume that P_α is minimal, then we have to assume that the "transformation matrix" $\mathbf{W}$ has full rank $m-1$, so has $m-1$ linearly independent rows. If not, first reduce P_α to a minimal family with dimension $t < |\mathscr{C}|$, then use a $(m-1) \times t$ dimensional matrix $\mathbf{W}$ has full rank $m-1$.

For example, consider the case with $\mathscr{S} = (|E|, |C_3|, |S_2|, |C_4|, |\triangleright\!-|, |C_5|)$ and $\alpha \in A = \mathbb{R}^6$. The approximating family has dimension 4, with sufficient statistics $W = \mathbf{W}\mathscr{S}$ equal to

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 & 0 \\ 2 & 0 & 4 & 8 & 0 & 0 \\ 0 & 30 & 0 & 0 & 10 & 10 \end{pmatrix} \mathscr{S} = \begin{pmatrix} 2|E| \\ 6|C_3| \\ 2|E| + 4|S_2| + 8|C_4| \\ 30|C_3| + 10|\triangleright\!-| + 10|C_5| \end{pmatrix}. \tag{5.9}$$

Also the approximating family is invariant under bijective linear transformation of W. Similarly to 5.5, every $\varphi_{\widetilde{\mathbf{B}},\tilde{b}} : \mathbb{R}^{m-1} \to \mathbb{R}^{m-1}$ defined as

$$\varphi_{\widetilde{\mathbf{B}},\tilde{b}}(W) = \widetilde{\mathbf{B}}W + \widetilde{b} \tag{5.10}$$

induces a model equivalent to Q_θ, if $\widetilde{\mathbf{B}}$ is nonsingular. Therefore, Q_θ can be interpreted as a low rank approximation of P_α, W is an interpretable sufficient statistics, and can be explicitly computed from G from its eigenvalues. All other $m-1$ dimensional families which are linear approximations of P_α, have sufficient statistics $\widetilde{\mathbf{B}}W$ and they are equivalent to Q_θ.

A reasonable choice is the transformation $\varphi_{\widetilde{\mathbf{B}},0}$ defined as

$$\widetilde{W} = \widetilde{\mathbf{B}}W = \begin{pmatrix} 1/2! & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/m! \end{pmatrix} W. \tag{5.11}$$

In this case $\widetilde{\mathbf{B}}$ standardize the walks in the same way as the Estrada index, defined in chapter 3. This transformation of the sufficient statistics is useful because the closed walks W does not "explode" when m is big. Moreover, the small walks are more important because they are formed only by small subgraphs. The correspondent parametrization is $\tilde{\theta}_k = k! \cdot \theta_k$, and the family $\widetilde{Q}_\theta$ with density proportional to $\exp(\tilde{\theta} \cdot \widetilde{W})$ is equivalent to Q_θ. So the two families represent the same information. When n is finite, as it has been assumed so far, also W_m is finite. Therefore

$$\widetilde{W}_m = \frac{W_m}{m!} \xrightarrow{m\to\infty} 0,$$

so the bigger sufficient statistics are heavily penalized. Since Q_θ and $\widetilde{Q}_\theta$ are equivalent, to keep the notation simpler in the chapter is always used Q_θ, despite the transformed family offers a better parametrization, in term of interpretation of the parameters.

The information loss depends on how large is m. If $m-1 = \mathbb{R}^{|\mathscr{C}|}$ and $\mathbf{W}$ has full rank, than it is invertible and Q_θ is equivalent to P_α. So if $|\mathscr{C}|$ is bounded, as in the case of a network model with fixed number of vertices, exists m large enough such that P_α and Q_θ are equivalent. The first family, because of the Hammersley-Clifford theorem, can represent every network model invariant under relabeling. Thus, exists m big enough such that the first m moments of the eigenvalue distributions contain all the information of an homogeneous network model.

The most interesting cases are when $m-1 < \mathbb{R}^{|\mathscr{C}|}$, or $m-1 \ll \mathbb{R}^{|\mathscr{C}|}$. Note that for every dependence assumption beyond Markov dependency, $|\mathscr{C}|$ is extremely large. However, it's reasonable to assume that smaller walks are more important than long ones, so we need to characterize the lost of information by a small rank approximation. The rows of $\mathbf{W}$ generate an $m-1$ dimensional affine subspace of $\mathbb{R}^{|\mathscr{C}|}$, so it is possible to find a $(|\mathscr{C}| - (m-1)) \times |\mathscr{C}|$ dimensional matrix $\mathbf{U}$ with full rank $|\mathscr{C}| - (m-1)$ such that

$$\mathbf{W}\mathbf{U}^T = \mathbf{0}, \tag{5.12}$$

where $\mathbf{0}$ is the matrix of zeros. The rows of $\mathbf{U}$ generate the affine subspace orthogonal to the one generated by $\mathbf{W}$, so a family equivalent to P_α has sufficient statistics

$$\widetilde{W} = \begin{pmatrix} W \\ U \end{pmatrix} = \begin{pmatrix} \mathbf{W} \\ \mathbf{U} \end{pmatrix} \mathscr{S}. \tag{5.13}$$

If P_α is approximated with Q_θ, the loss of information is U. This information is in the orthogonal complement of the space spanned by the rows of $\mathbf{W}$.

The decomposition of $\mathbb{R}^{|\mathscr{C}|}$ in the two orthogonal affine subspaces spanned by $\mathbf{W}$ and $\mathbf{U}$ implies a similar decomposition of the parametric space. The families P_α and Q_θ have parametric spaces $A = \mathbb{R}^{|\mathscr{C}|}$ and $\Theta = \mathbb{R}^{m-1}$ respectively. The approximating family is parametrized by $\theta \in \Theta = \mathbb{R}^{m-1}$. The relation 5.12 implies that

$$\mathbf{U}\alpha = \mathbf{U}\alpha_0 \quad \Longleftrightarrow \quad \exists \theta \in \mathbb{R}^{m-1} : \ \alpha = \alpha_0 + \mathbf{W}^T\theta. \tag{5.14}$$

Then

$$\left\{\alpha \in \mathbb{R}^{|\mathscr{C}|} : \mathbf{U}\alpha = \mathbf{U}\alpha_0\right\} = \left\{\alpha_0 + \mathbf{W}^T\theta, \theta \in \mathbb{R}^{m-1}\right\}, \tag{5.15}$$

(Jørgensen and Labouriau (2012)). The real low rank family is denoted as $P_{\alpha_0 + \mathbf{W}^T\theta}$, and it's called *affine hypothesis* for P_α.

The decomposition of the original space in two orthogonal subspaces is es-

pecially useful for exponential families. The original family can be written as

$$\begin{aligned} P_\alpha(G) &= e^{\alpha_0\cdot\mathscr{S}+(\mathbf{W}^T\theta)\cdot\mathscr{S}-\psi(\alpha_0+\mathbf{W}^T\theta)} = e^{\alpha_0\cdot\mathscr{S}+\theta\cdot(\mathbf{W}\mathscr{S})-\psi(\alpha_0+\mathbf{W}^T\theta)} = \\ &= e^{\alpha_0\cdot\mathscr{S}+\theta\cdot W-\psi(\alpha_0+\mathbf{W}^T\theta)}. \end{aligned} \tag{5.16}$$

If $\alpha_0 = 0$, then $P_\alpha(G) = Q_\theta(G) = e^{\theta\cdot W-\phi(\theta)}$ and the relation between the partition functions of the real and approximating family is explicit:

$$\phi(\theta) = \psi(\mathbf{W}^T\theta). \tag{5.17}$$

In this case the parameters of the original family P_α are in the linear subspace spanned by the rows of $\mathbf{W}$. Therefore the minimal sufficient statistics are $\mathbf{W}\mathscr{S}$ and Q_θ is obtained from reduction to sufficiency.

Information Projection

So far it has been exposed the decomposition of the model in the approximating family and information loss. When the last one is zero, the real and approximating families are equivalent. However this is usually not the case, so we want to choose a specific distribution in the approximating family, which is the closest to the real distribution. This is the information projection introduced in chapter 2, it is the distribution that minimizes the approximation loss between the real and all distributions in the approximating family. Differently than the density estimation problem in chapter 2, for random graphs the relative entropy can be computed explicitly, and depends on the algebraic decomposition of A and $\mathscr{S}$ outlined above.

For a fixed $\alpha^* \in A$, we can find the distribution in the family Q_θ which is closer to $P^* = P_{\alpha^*}$, the real distribution of the network. The solution is $Q^* = Q_{\theta^*}$, and it is an approximation of the real distribution (Q^* can be interpreted as an estimator of P^* in information sense). Q^* is the information projection of the family Q_θ to the set of distributions with same moments than P^*_α. More precisely consider the following

$$\begin{aligned} \mu_{\mathscr{S}} &= \mathbb{E}_{P^*}(\mathscr{S}) = \nabla_\alpha\psi(\alpha)\big|_{\alpha=\alpha^*} \\ \mu_W &= \mathbb{E}_{P^*}(W) = \mathbf{W}\mathbb{E}_{P^*}(\mathscr{S}) = \mathbf{W}\mu_{\mathscr{S}} \\ \eta_W(\theta) &= \mathbb{E}_{Q_\theta}(W) = \nabla_\theta\phi(\theta) \end{aligned} \tag{5.18}$$

The information projection is $Q^* = Q_{\theta^*}$, where θ^* is the only solution of

$$\eta_W(\theta) = \mu_W \quad \Longleftrightarrow \quad \theta^* = \eta_W^{-1}(\mathbf{W}\mu_{\mathscr{S}}). \tag{5.19}$$

The relation between θ^* and the approximation error is

$$\theta^* = \operatorname*{argmin}_{\theta\in\Theta} D(P^*||Q_\theta), \tag{5.20}$$

so for a fixed α^*, Q^*_θ is the distribution in Q_θ closest to P^*. The minimum exist

and is unique by theorem 3.1 in Csiszár (1975).

Using 5.16 and 5.17, the approximation error can be written explicitly as

$$\begin{aligned} D(P^*||Q^*) &= D(P_{\alpha_0^*+\mathbf{W}^T\theta^*}||Q_{\theta^*}) = \\ &= \alpha_0^* \cdot \mathbb{E}_{P^*}(\mathscr{S}) - \left(\psi(\alpha_0^* + \mathbf{W}^T\theta^*) - \phi(\theta^*)\right) = \\ &= \alpha_0^* \cdot \mathbb{E}_{P^*}(\mathscr{S}) - \left(\psi(\alpha_0^* + \mathbf{W}^T\theta^*) - \psi(\mathbf{W}^T\theta^*)\right). \end{aligned} \tag{5.21}$$

Then, the difference of partition functions can be expanded using Taylor series:

$$\alpha_0^* \cdot \psi^{(1)}(\mathbf{W}^T\theta^*) + \frac{1}{2}\alpha_0^* \cdot \psi^{(2)}(\mathbf{W}^T\theta^*)\alpha_0^* + \dots , \tag{5.22}$$

where

$$\begin{aligned} \psi^{(1)}(\mathbf{W}^T\theta^*) &= \nabla_\alpha \psi(\alpha)\big|_{\alpha=\mathbf{W}^T\theta^*} = \mathbb{E}_{P_{\mathbf{W}^T\theta^*}}(\mathscr{S}), \\ \psi^{(2)}(\mathbf{W}^T\theta^*) &= \mathrm{H}_\alpha \psi(\alpha)\big|_{\alpha=\mathbf{W}^T\theta^*} = \mathbb{V}_{P_{\mathbf{W}^T\theta^*}}(\mathscr{S}). \end{aligned} \tag{5.23}$$

The Hessian is positive definite because the log-partition function ϕ is convex. Therefore if α_0^* is close to 0, the approximation loss is

$$D(P^*||Q^*) \simeq \alpha_0^* \cdot (\mathbb{E}_{\alpha^*}(\mathscr{S}) - \mathbb{E}_{\mathbf{W}^T\theta^*}(\mathscr{S})) - \frac{1}{2}\alpha_0^* \cdot \mathbb{V}_{\mathbf{W}^T\theta^*}(\mathscr{S})\alpha_0^*. \tag{5.24}$$

5.2 ERGM as Parametrized Centrality Measure

One of the main descriptive statistics of network data are the centrality measures, which rank the vertices based on their importance in the structure of the network. Markov random graphs parametrize the degree distribution, while the method introduced in the previous section parametrizes the moments of the eigenvalue distribution. Therefore the exponential random graph models induces a parametrized centrality measure.

The degree centrality introduced in section 3.2 assigns importance to the vertices based on the size of the neighbourhoods. Markov random graphs can be written with the family of distributions

$$P_{\tau,\mu}(G) = e^{\tau|C_3| + \sum_{i=0}^{n-1} \mu_i \tilde{d}_i - \widetilde{D}(\tau,\mu)}. \tag{5.25}$$

Therefore the *parametrized degree centrality* is

$$d_\mu = (\mu_{(1)} \cdot d_1, ..., \mu_{(n)} \cdot d_n), \tag{5.26}$$

where $\mu_{(i)}$ is the parameter associated to $\tilde{d}_{(i)}$, the proportion of vertices with degree equal to d_i. Therefore, the parameter μ rescale the importance of the size of the neighbourhoods. The centrality d_μ as it is looks too flexible, this because the original Markov graph $P_{\tau,\mu}$ is indeed too flexible because it's degenerate for most values for τ and μ. However, the centrality 5.26 can be easily defined for a non-degenerate curved model as $d_{\mu(\eta)}$, where η is lower dimensional.

In equation 3.18 the Estrada or subgraph centrality has been defined as

$$\widetilde{EE} = \mathrm{diag}(e^{\mathbf{A}}) = \mathrm{diag}(n) + \sum_{k=2}^{\infty} \frac{1}{k!}\mathrm{diag}(\mathbf{A}^k). \tag{5.27}$$

This measure can be extended as

$$\widetilde{EE}_{\theta^*,m} = \mathrm{diag}(n) + \sum_{k=2}^{m} \frac{\theta_k^*}{k!}\mathrm{diag}(\mathbf{A}^k), \tag{5.28}$$

which is the diagonal of a parametrized matrix exponential function, and will be denoted as *parametrized subgraph centrality*. The last equation is equivalent to

$$\widetilde{EE}_{\theta^*,m} = \mathrm{diag}(n) + \mathrm{diag}\left(\mathbf{\Gamma}\left(\sum_{k=2}^{m} \frac{\theta_k^*}{k!}\sum_{i=1}^{n}\lambda_i^k\right)\mathbf{\Gamma}^\top\right), \tag{5.29}$$

where $\mathbf{\Gamma}$ is the matrix of eigenvectors of $\mathbf{A}$ and $\lambda_1, ..., \lambda_n$ are its eigenvalues. The Estrada centrality is $\widetilde{EE}_{\theta^*,m}$ where $\theta^* = 1$ and $m = \infty$. When m diverges, the approximating family is equivalent to the original, therefore in this case the parametrized subgraph centrality ranks its vertices using all "homogeneous information" contained in the network.

5.3 Final Remarks

The approximating model $Q_{\theta,m}$ can be used to approximate every homogeneous distribution P_α. However, the approximation error has been evaluated only in term of information, but both models are useful only when they are non-degenerate. This section is a short discussion on the next steps of the research.

$Q_{\theta,m}$ has a natural interpretation in term of closed walks of the network, therefore uses particularly the information of the tie placement at local level. Counts of closed walk of different length are associated with moments of the eigenvalue distribution of the network. The first moments are explicit functions of small subgraphs. The number of sufficient statistics m included in the Hamiltonian can be related with how much the properties of the generative process are reflected in the local behaviour. When m grows all the information is used and the exponential random graph model can approximate arbitrarily well an homogeneous network.

However, P_α and $Q_{\theta,m}$ have been introduced in canonical form, therefore both models are unstable. The approximation error, in term of information can be written explicitly, in practise have to be evaluated only for non-degenerate models. Therefore, a future research question will the study of

$$\begin{aligned} D(P^*||Q^*) &= \alpha_0^* \cdot \mathbb{E}_{P^*}(\mathscr{S}) - \big(\psi(\alpha_0^* + \mathbf{W}^T\theta^*) - \psi(\mathbf{W}^T\theta^*)\big) \\ &\simeq \alpha_0^* \cdot (\mathbb{E}_{\alpha^*}(\mathscr{S}) - \mathbb{E}_{\mathbf{W}^T\theta^*}(\mathscr{S})) - \alpha_0^* \cdot \mathbb{V}_{\mathbf{W}^T\theta^*}(\mathscr{S})\alpha_0^*/2, \end{aligned} \tag{5.30}$$

for two stable curved parametrizations $\alpha^* = \alpha(\eta^*)$ and $\theta^* = \theta(\beta^*)$, where η^* and β^* are low dimensional parameters.

The approximating ERGM, contains as sufficient statistics the first m moments of the eigenvalue distribution. The eigenvalues does not contain information on non-homogeneous structures of the network, like communities for example, as this information is only in its eigenvectors. Therefore, Q_θ cannot be used to model inhomogeneous networks. Despite so, we can choose which sufficient statistics include in the model, so it can be used a "combined" Hamiltonian

$$\sum_{k=4}^{m} \theta_k W_k + \left(\sum_{j \in J_3} \tau_j |C_{3;j}| + \sum_{i=1}^{n-1} \sum_{j \in J_i} \alpha_{i;j} |S_{i;j}| \right), \tag{5.31}$$

where

$$J_i = \{(j_1, ..., j_i) \text{ s.t. } j_l \in \{1, ..., C\}\}. \tag{5.32}$$

The second part of equation 5.31 is the Hamiltonian of a Markov graph with vertices in C communities, the first part models homogeneously the residual non-Markov dependency.

Bibliography

Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

Andrew R Barron and Chyong-Hwa Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, pages 1347–1369, 1991.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

Lawrence D Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-monograph series*, 9: i–279, 1986.

Maurizio Cailotto. *Algebra e Geometria Lineari e Quadratiche*. 2004.

Stéphane Canu and Alex Smola. Kernel methods and the exponential family. *Neurocomputing*, 69(7-9):714–720, 2006.

Arun G Chandrasekhar and Matthew O Jackson. Tractable and consistent random graph models. Technical report, National Bureau of Economic Research, 2014.

Sourav Chatterjee, Persi Diaconis, et al. Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461, 2013.

Bradford R Crain. An information theoretic approach to approximating a probability distribution. *SIAM Journal on Applied Mathematics*, 32(2):339–346, 1977.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

Bradley Efron et al. The geometry of exponential families. *The Annals of Statistics*, 6(2):362–376, 1978.

Paul Erdös and Alfréd Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

Ernesto Estrada. *The structure of complex networks: theory and applications*. Oxford University Press, 2012.

Ernesto Estrada and Naomichi Hatano. Statistical-mechanical approach to subgraph centrality in complex networks. *Chemical Physics Letters*, 439(1-3): 247–251, 2007.

Ove Frank and David Strauss. Markov graphs. *Journal of the american Statistical association*, 81(395):832–842, 1986.

Charles J Geyer and Elizabeth A Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 657–699, 1992.

Adom Giffin. Maximum entropy: The universal method for inference. *arXiv preprint arXiv:0901.2987*, 2009.

Mark S Handcock, Garry Robins, Tom Snijders, Jim Moody, and Julian Besag. Assessing degeneracy in statistical models of social networks. Technical report, Citeseer, 2003.

Mark S Handcock, David R Hunter, Carter T Butts, Steven M Goodreau, and Martina Morris. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of statistical software*, 24 (1):1548, 2008.

K. Huang. *Statistical mechanics*. Wiley, 1963. URL `https://books.google.nl/books?id=MolRAAAAMAAJ`.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

Bent Jørgensen and Rodrigo S Labouriau. Exponential families and theoretical inference. 2012.

László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

Nathaniel FG Martin and James W England. *Mathematical theory of entropy*, volume 12. Cambridge university press, 2011.

Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

Mark Newman. *Networks: an introduction*. Oxford university press, 2010.

Steve Pressé, Kingshuk Ghosh, Julian Lee, and Ken A Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics*, 85(3):1115, 2013.

Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

Michael Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496): 1361–1370, 2011.

Cosma Rohilla Shalizi and Alessandro Rinaldo. Consistency under sampling of exponential random graph models. *Annals of statistics*, 41(2):508, 2013.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

John Shore and Rodney Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26(1):26–37, 1980.

Tom AB Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.

Tom AB Snijders, Philippa E Pattison, Garry L Robins, and Mark S Handcock. New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153, 2006.

David Strauss and Michael Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990.

Johannes Van Der Pol. Introduction to network modeling using Exponential Random Graph models (ERGM). working paper or preprint, October 2017. URL `https://hal.archives-ouvertes.fr/hal-01284994`.

Piet Van Mieghem. *Graph spectra for complex networks*. Cambridge University Press, 2010.