Journal de la Société Française de Statistique
**Vol. 156** No. 3 (2015)

# Random graph models: an overview of modeling approaches

**Titre:** Modèles de graphes aléatoires : un panorama des démarches de modélisation

Antoine Channarond[1]

**Abstract:** This article nonexhaustively reviews random graph models designed to model interaction networks. It begins with the Erdős-Rényi model. It has been deeply studied, as it is based on simple assumptions: independence and homogeneity of the links, which are however too simplistic for applications. The article then focuses on modeling approaches of the hetereogeneity and of the dependences between the links. It starts from probabilistic models reproducing generative processes of the real-world networks (Barabási-Albert or Watts-Strogatz models for instance) and arrives to models more suitable for statistics. Exponential models (ERGM or $p^*$) enable to introduce dependences between the desired links. Models with latent variables enable to model heterogeneity of the population and to analyze it.

**Résumé :** Cet article établit une revue non exhaustive des modèles de graphes aléatoires destinés à la modélisation de réseaux d'interaction. Il commence par le modèle d'Erdős-Rényi qui a pu être étudié en profondeur car il repose sur des hypothèses simples d'indépendance et d'homogénéité des liens, cependant trop réductrices pour les applications. L'article se concentre ensuite sur les démarches de modélisation de l'hétérogénéité et des dépendances entre les liens. Il part de modèles probabilistes reproduisant les processus de génération des réseaux réels (modèles de Barabási-Albert ou de Watts-Strogatz par exemple) et arrive à des modèles plus adaptés à la statistique. Les modèles exponentiels (ERGM ou $p^*$) permettent d'introduire des dépendances entre les liens voulus. Les modèles à variables latentes permettent de modéliser l'hétérogénéité de la population et de l'analyser.

***Keywords:*** random graph models, review, Erdős-Rényi model, complex networks
***Mots-clés :*** modC(les de graphes aléatoires, revue, modC(le d'Erdős-Rényi, réseaux complexes
***AMS 2000 subject classifications:*** 05C80, 05C82, 90B15, 91D30, 62-02, 60-02

Whatever the topic is, sociology, biology or computer science, understanding group phenomena requires to collect and analyze data not only on the members of the group but on their interactions as well. The set of the interactions within a group forms what is called an interaction network. The network research field is at a crossroads of many and varied scientific fields, and some of them can be brothers from the point of view of networks:

- An epidemy can be formalized in the same framework as the propagation of a rumor through a group of people (Draief and Massoulié, 2010). The common issue is to predict whether the disease (the rumor) will infect the whole population or a negligible part, via the study of the network properties.
- The study of the vulnerability of networks can be useful to prevent computer networks from disruptions by random breakdowns or targeted attacks (Bollobás and Riordan, 2004a), as well as farm networks from loss of seeds because of storms (Barbillon et al., 2015).

[1] UMR6085 Laboratoire de Mathématiques Salem, Université de Rouen.
E-mail: antoine.channarond@univ-rouen.fr


 ISSN: 2102-6238

– Structure of protein-protein interaction networks and social networks can be analyzed in an unified statistical framework: the regulating action of the proteins is equivalent to the social role of people (Picard et al., 2009; Snijders and Nowicki, 1997).

**Roots in sociology.** Social networks are probably the most popular example of this concept, and the network research field is deeply rooted in sociology. In the late nineteenth century, the father of the modern sociology, Émile Durkheim, explained that interactions within a population shape its global behaviour, and individual factors are not sufficient (Durkheim, 1893). Then in the thirties, while sociometry emerged from the desire to arm sociological theories with mathematical arguments (Moreno, 1937), the sociologist Moreno began to represent graphically the relations between people with diagrams of points and lines called sociograms, and thus started to formalize the concept of social network (Moreno and Jennings, 1938).

**Origins of graphs.** Graph theory was initiated in the eighteenth century and became popular with Euler's elegant solution of the problem of the seven bridges of KC6nigsberg using graphs (Euler, 1741). Then they stayed for a long time nice objects rather designed for algebraic or topological issues [1], until they were finally thought of as a mathematical representation of networks in the fourties, for example in articles of the mathematician and social scientist Luce (1952), still in a deterministic setting.

**What type of graphs are considered.** Graphs are mathematical objects modeling interaction networks. They are basically composed of a set $V$ representing the group of individuals (called nodes or vertices), and of the set $E$ of the interactions (called edges or links). The type of the interaction may depend on the application. This paper deals only with pairwise, binary and reciprocal interactions, meaning that: interactions are assumed to involve exactly two individuals — in particular there is no self-interaction —, only the presence or absence of interaction is considered, not its intensity (see weighted graphs for this variant), and interactions are not directed: for any $i, j \in V$, $i$ interacts with $j$ implies that $j$ also interacts with $i$ (otherwise, see directed graphs to distinguish between a direction and the other one). In this setting, interactions can be represented by unordered pairs of elements of $V$; if $i, j \in V$, then $\{i, j\} \in E$ if individuals $i$ and $j$ interact. Data is commonly in the form of an adjacency matrix $X$, defining a graph in the following way: for all $i, j \in V$, $X_{ij} = 1$ if $i$ and $j$ interact, or 0 if they do not. See Section 1 for more details.

**Erdős-Rényi model.** Discrete probability and computer science introduced the first random graph models in 1959 when the founding works of Erdős and Rényi (1959) and Gilbert (1959) were published. Their models are based on two very nice assumptions, the independence and the homogeneity of the links: any two individuals are connected with probability $p$. The first part of this paper addresses this model and some of its features. Although the assumptions of the model allow its exhaustive study, they are actually too simplistic and graphs drawn from these models do not satisfy features of empirical networks. The paper explains how it can be used nevertheless, as a tool to study other models or as model of a null hypothesis in statistical tests.

[1] See works of Cayley, Sylvester or Kirchhoff in the nineteenth century for examples.


 ISSN: 2102-6238

**First heterogeneous models.** Two of the major focuses in the network research field are precisely how to model the heterogeneity and the dependences between the links. Social sciences and statistics began to join forces from the early eighties to produce heterogeneous models. They often derived from sociological concepts established in the seventies: structural equivalence (Lorrain and White, 1971), social distance (McFarland and Brown, 1973) for examples, themselves based on many empirical studies from the sixties in sociology. Holland and Leinhardt (1981) introduced the $p_1$ model, taking directly social characteristics of the individuals into account simply via logistic regression; Holland et al. (1983) initiated block models, deriving from the concept of structural equivalence. Besides models from the exponential family, once adapted to graphs (also known as $p^*$ models), allowed to add dependences between links in a simple fashion; for example between links sharing one common node (Frank and Strauss, 1986) (also known as Markov random graph models). However beyond this case, these models become quickly untractable as soon as too many dependences are introduced. Thus a trade-off is desirable between on the one hand the feasibility of the study of the model and on the other hand its goodness of fit to real-world networks.

**Dependences in the specific data configuration of graphs.** The least complexification of the models makes their study and therefore their inference much harder. Indeed interaction data is an unusual framework, as each variable $X_{ij}$ of the data involves two individuals instead of one. This configuration may create extra complex dependences, but may also be a strength, as information on each individual grows with the number of individuals (see conclusion of 7.5). Specific attention is paid in this paper to the dependence structure of the statistical models. Developing efficient inference methods in dependent contexts is still a very active field today, which is fruitful in particular via approaches of approximate inference like variational approximations or pseudo-likelihoods for instance.

**Era of the big data.** The number and size of network data have gone up since the beginning of network analysis, from some tens of nodes at that time, to millions and billions today. The first explosion of the data volume date back to the massive digitization of the data in the nineties. A stronger aftershock arose at the start of the twenty-first century, based on the democratization of the Internet which led to the emergence of many and varied large networks: physical infrastructures, the World Wide Web, and a plethora of online social and sharing networks above all. In parallel, as sequencing of genomes became current, biologists got interested in genetic interaction networks, which are quite large as well. These recent revolutions created two veins in the network research field in the late nineties. Analyzing these new large graph datasets within a reasonable amount of time became another challenging issue. Interaction data grows very fast as function of the number of individuals and its analysis brings combinatorial issues. Thus new inference methods need to be efficient from a computational point of view.

**Emergence of universal features.** Moreover, a large part of the literature focused on universal features of the real-world networks, which appeared more obviously in these new numerous huge graph datasets by a law of large numbers effect. In particular physics and computer science proposed hypotheses on the way how networks grow or are generated, to explain how these universal features emerge, for exemple the scale-free property (Barabási and Albert, 1999) or the


 ISSN: 2102-6238

small-world property (Watts and Strogatz, 1998). Probabilistic models based on these assumptions were constructed and some of them are briefly presented in this paper.

**Models with latent variables.** Since the 2000s much attention is paid to models with latent variables in sociology, biology jointly with statistics, since they allow not only to model the heterogeneity underlying the network, but also to analyze it via model-based clustering. The literature about the Stochastic Blockmodel (Snijders and Nowicki, 1997) or its variants and about graphons (Lovász and Szegedy, 2006) is especially prolific at that time. These models are carefully described in this paper.

**Organization and goal.** The paper provides an insight into random graphs models and their features, focusing on how to model heterogeneity and dependences between the links. It also reviews main results characteristic of these models, or typical methods for their inference, regarding the statistical models. The topic is vast, the literature burgeoning, and the present paper is of course not exhaustive at all. To go further, see for example reviews of Goldenberg et al. (2010); Kolaczyk (2009); Snijders (2011); Matias and Robin (2014); Gerbaud (2010). The present article is organized as function of the purposes models have been designed for. It distinguishes between three main modeling approaches, each part of the paper corresponding to one of these.

The first approach is illustrative and is typical of the toy model, embodied by the Erdős-Rényi model. It has been deeply studied in the literature and helps to figure out basic notions and what is at stake in the field. After some definitions used in the whole paper (Section 1), the first part presents Erdős-Rényi model (Section 2) and how it can be interpreted to model the propagation of a rumor (or an epidemy) with the Reed-Frost model (Subsection 2.2). Then it briefly outlines main features of the Erdős-Rényi model, and emphasizes that their emergence has generally a phase transition with respect to the probability $p$ of connection (Subsections 2.3 and 2.4). Basic assumptions of the model, independence and homogeneity, are also criticized in Subsection 2.5, highlighting what is not plausible from the point of view of the application.

The second part (Sections 3 and 4) firstly describes some of the key-features which are shared by most real-world networks (Section 3), and are therefore expected from random graph models with high probability. The second approach achieved by the models then introduced in this part, is to satisfy some of these key-features, by mimicking the way the nature generates heterogeneous networks. Each of these models has its own growth dynamics or generating process, proposed as an hypothesis explaining the emergence of a target feature. For example, small-world models try to account for the low diameter of social networks with mechanisms producing shortcuts through the graph (Subsection 4.1), and Barabási-Albert's model attempts to explain emergence of hubs and the heavy-tailed degree distribution with the preferential attachment (Subsection 4.3). Random geometric graphs are also briefly mentioned and are helpful to model neighborhood networks (Subsection 4.2). Even though these models successfully achieve their objective, they are not designed for statistical inference and they can only poorly fit to observed data, as the small number of parameters makes them too inflexible.

The third approach is precisely at the heart of statistics and central to this paper: the third part (Sections 5 to 7) of the paper is hence more detailed. Whereas models from the previous part have to be not too complex to allow mathematical proofs of their target properties, models designed for statistics have to achieve one more trade-off: they are expected to allow correct


 ISSN: 2102-6238

fitting to observed data and efficient methods of inference in addition. Logistic regression models (an undirected variant of the $p_1$ model) are briefly illustrated through an introducing example in Section 5. Links are still independent, and characteristics of the individuals accounting for heterogeneity are assumed to be known. Two large families of models are then presented in this part. Firstly exponential random graph models (ERGM, or $p^*$ models) are described in the most general setting in Section 6. They enable to introduce dependences between links by controlling occurence probabilities of some link configurations (or motifs) of interest. It especially focuses on Markov random graph models (Subsection 6.1), which add dependences between links sharing one common node by controlling stars and triangle motifs. Some methods of estimation of the parameters are also addressed, and issues related to dependences are evoked in Subsection 6.3. $p_1$ and $p^*$ models are heterogeneous, but they do not allow the analysis of the heterogeneity, whereas models with latent variables do. These models are introduced in a unified framework in Subsection 7.2 and some special cases are much more detailed: the Stochastic Blockmodel (Subsection 7.4), latent position models (Subsection 7.6) and graphons (Subsection 7.8). Inference methods are evoked for all of them, with a particular emphasis on the variational approximation for the stochastic blockmodel in Subsection 7.5. Issues related to the complex dependence structure of such models are described in detail, and motivate the use of such an approximation. Moreover unlike previous models, these allow to analyze the heterogeneous structure underlying the observed networks, and in particular, to cluster the individuals. The model has to be conveniently chosen among this family because clusters strongly depend on this choice. A discussion in 7.7 addresses this question.

**What is not addressed in this paper.** Note that this paper does not cover topics like inference of networks, insofar as networks is the observed data and is not to infer (Charbonnier et al., 2010); Bayesian networks or graphical models and their inference (Nielsen and Jensen, 2009; Murphy, 2002), even though the directed acyclic graph of the models with latent variables are evoked in Section 7.5; computational issues of quantities related to graphs like diameter, shortest paths, clustering coefficient, etc. We are just interested in how they behave in some of the reviewed random graph models. Non-model-based clustering methods are not in the scope of the paper, see Fortunato (2010) for a review. Clustering is nevertheless mentioned because it is allowed by models with latent variables but is not a goal per se. Percolation on (random) graphs (Callaway et al., 2000), pure probabilistic theory on random graphs (Bollobás, 2001; Van Der Hofstad, 2009; Durrett, 2007) are also not addressed. This list is of course not exhaustive.

## 1. Some definitions

According to the type of interactions involved in the group of interest, the graph as a mathematical object may have varied definitions. In this paper we are modeling the presence or the absence of an interaction between the individuals, and not the number of interactions they have. Moreover an individual is not supposed to interact with itself and interactions are assumed to be reciprocal.

Let us define the object matching with these assumptions. A simple undirected graph is a couple $(V, E)$ where $V$ is a set and $E$ a set of pairs of elements from $V$. Elements of $V$ are called nodes (or vertices) and those of $E$ are called edges. Note that since edges are represented by a set of node pairs, this definition rules out self-loops (edges going from a node to itself), multiple


 ISSN: 2102-6238

edges and asymmetric edges. A graph with no self-loop and no multiple edges is said to be simple, and a graph with only symmetric edges is said to be undirected.

A simple undirected graph $(V,E)$ is said to be finite if $V$ is a finite set. If $i,j \in V$ such that $i \neq j$ and $\{i,j\} \in E$, $i$ and $j$ are said to be connected or neighbours. For all integer $n \geq 1$, the set of the integers between 1 and $n$ is denoted by $[n]$. The size of a finite simple graph is defined as the cardinality of $V$. To simplify, for all graphs with size $n$, we generally use $V = [n]$ in this paper.

**Subgraphs and induction.** Any graph $(V',E')$ is said to be included in the graph $(V,E)$, or is a subgraph of $(V,E)$ if $V' \subset V$ and $E' \subset E$. The subgraph $(V',E')$ induced by $I \subset V$ is the graph $(I, E \cap \mathscr{P}(I))$, where $\mathscr{P}(I)$ is the set of all pairs of elements of $I$. Define also $\mathscr{P}_n = \mathscr{P}([n])$.

A clique is a subgraph such that each node is connected to all other nodes. A triangle is a clique composed of three nodes.

**Adjacency matrix.** $(V,E)$ is a finite undirected simple graph with size $n$, its adjacency matrix is the matrix $x \in \{0,1\}^{n \times n}$ such that for all $i,j \in [n]$, $x_{ij} = \mathbb{1}_{\{i,j\} \in E}$, i.e. $x_{ij} = 1$ if and only if $i$ and $j$ are connected in $(V,E)$, 0 otherwise. A graph is completely described by its adjacency matrix, so a graph and its adjacency matrix will be always identified. As graphs considered in this paper are undirected, their adjacency matrices are symmetric. Moreover, their diagonals are zero, since the graphs are simple and hence have no self-loop.

**Degree.** Let $x$ be the adjacency matrix of a graph $g$ with size $n$. The degree of the node $i \in [n]$ is defined as the number of neighbours of $i$ in $g$. It is denoted by $D_i^g$ or just $D_i$ if the context is clear and equals:

$$D_i^g = \sum_{j=1}^{n} x_{ij}.$$

**Paths and connectedness.** Let $i,j \in [n]$. A path between $i$ and $j$ in the graph $g$ with size $n$ is a finite sequence $i_0, i_1, \ldots, i_r \in [n]$ such that $i_0 = i$, $i_r = j$ and for all $1 \leq k \leq r$, $i_{k-1}$ and $i_k$ are connected in $g$. The length of the path is the integer $r$. This defines an equivalence relation on the node set. The equivalence classes of this equivalence relation are called connected components. If a graph has only one connected component, it is said to be connected.

Paths allow to define a concept of distance in the graph: the distance between two nodes is defined as the length of the shortest path between them. A cycle is a path whose length is non-zero, which arrives to the same node it started from. A tree is a graph without any cycle.

**Density.** We call density of the graph $g$ with size $n$ and adjacency matrix $x$, the proportion of present edges over the number of possible edges. It is denoted by $\zeta(g)$ and therefore equals:

$$\zeta(g) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} x_{ij}.$$

**Random graphs.** A random graph with size $n \in \mathbb{N}^*$ is here a random variable taking its values in the set of simple undirected graphs with $n$ nodes. Such a random graph is here directly defined by their adjacency matrix, generally denoted by $X = (X_{ij})_{i,j \in [n]}$. It is therefore a symmetric random matrix taking values in $\{0,1\}^{n \times n}$, and whose diagonal is zero. For all $i,j \in [n]$, $X_{ij}$ is therefore a


 ISSN: 2102-6238

Bernoulli distributed variable. A random graph model is a collection of distributions on graphs or equivalently on such matrices.

## 2. Erdős-Rényi model as tool and mathematical reference

### 2.1. *Definitions and first features of the model*

In the past, two models have been actually associated with the name of Erdős-Rényi, the uniform model (Erdős and Rényi, 1959) with a fixed number $M \in \mathbb{N}^*$ of edges, denoted by $\mathscr{G}(n,M)$, and the binomial model (Gilbert, 1959), denoted by $\mathscr{G}(n,p)$ where $p \in [0,1]$ is the connection probability between any two nodes, and is fixed. Today the name Erdős-Rényi is rather given to the binomial model. These models are not formally equivalent, but their features are very similar to each other in some asymptotic framework. They have been studied in all their guises, the second one more intensively. The reader can find many results in Bollobás (2001), Janson et al. (2000) or in Durrett (2007).

**Uniform model (Erdős and Rényi, 1959)** In this model, the random graph $X$ is uniformly drawn from the set of graphs with $n$ nodes and $M$ edges. It is sometimes called " the " random graph precisely insofar as the network topology is random and uniformly drawn. The choice of such a graph can be made by choosing $M$ pairs of nodes which will be the edges; note that there are $N = \binom{n}{2}$ possible pairs in a graph with $n$ nodes. Then the number of graphs with $n$ nodes and $M$ edges is $\binom{N}{M}$. Thus for all graph $x$ with $n$ nodes and $M$ nodes,

$$P_{\mathscr{G}(n,M)}(X = x) = \binom{N}{M}^{-1}.$$

The probability that an edge is present is the same for all node pairs and equals $M/N$. This model was introduced by Paul Erdős and Alfréd Rényi in Erdős and Rényi (1959) and was the subject of several articles, for example Erdos and Rényi (1961); Erdős and Renyi (1961). The theoretical model has been very well studied from the point of view of theoretical probability, but it is still rather unpractical to use, since the edge number is generally not deterministic in the applications.

**Binomial model (Gilbert, 1959)** In this model, a sequence of mutually independant Bernoulli distributed variables $(X_{ij})_{1\leq i<j\leq n}$ is drawn, with the same parameter $p \in ]0,1[$ for all of them. The random graph is represented by the symmetric adjacency matrix $X = (X_{ij})_{i,j\in[n]}$. $X$ has a random number of edges, which is binomially distributed, with parameters $(N,p)$, where $N$ still equals $\binom{n}{2}$. All graph with $n$ edges and $M$ edges is drawn with probability $p^M(1-p)^{N-M}$. The distribution of $X$ conditionally on the fact that $X$ has got $M$ edges is then actually $\mathscr{G}(n,M)$. Finally, the degrees are binomially distributed. For all $i \in [n]$,

$$D_i^X \sim \mathscr{B}(n-1,p).$$

The model $\mathscr{G}(n,p)$, already studied by Erdős at the end of the 40's, was also introduced in Gilbert (1959).


 ISSN: 2102-6238

**Link between the models** Let us consider the particular asymptotic framework in which the edge number $M$ for the first model and the connection probability $p$ for the second one depends on the graph size $n$. The edge number in $\mathscr{G}(n,p)$ is binomially distributed; by Hoeffding's inequality it implies that this variable concentrates very fast around its mean $Np$. If $p_n = M/N$, $\mathscr{G}(n,p_n)$ thus "behaves" the same way as $\mathscr{G}(n,M)$. Janson et al. (2000) clarifies this: in few words, most of graph properties are satisfied in $\mathscr{G}(n,p_n)$ asymptotically almost surely (a.a.s.) if and only if they are satisfied a.a.s. in $\mathscr{G}(n,M)$.

### 2.2. *Example of application of the Erdős-Rényi model*

Now we are presenting an example of application of the Erdős-Rényi model in epidemiology, so as to illustrate how to model a problem with that model and which practical modeling assumptions are at stake in it, as well as to motivate the study of its properties presented in the next sections. Let us introduce the Reed-Frost model and show the equivalence with Erdős-Rényi.

The Reed-Frost model [2] belongs to the class of epidemic models called SIR (for Susceptible, Infected, Recovered). In these models, each individual of the population is in one of the states S, I or R, depending on whether it has never been infected by the disease and *susceptible* to be, or is currently *infected*, or *recovered*, that is, immune. The Reed-Frost model is a discrete version of this kind of models. Only one individual is infected at the time origin. The dynamics is then as follows:

1. The population is closed: the disease can be transmitted only within the $n$ individuals.
2. The disease can be transmitted through only one kind of contact and the probability $p$ to have this contact in a time unit is the same between any two individuals and at any time unit.
3. Contacts between individuals are independant.
4. An infected individual remains so only for a time unit and is then definitively immune.

Barbour and Mollison (1990) first introduced the mathematical construction of the Reed-Frost model from the Erdős-Rényi model. $X$ is assumed to be a graph drawn under the model $\mathscr{G}(n,p)$. Let $i \in [n]$ be the individual infected at the origin. $\delta_X$ denotes the distance in the graph $X$. The state $e_t(j)$ of the individual $j \in [n]$ at time $t \in \mathbb{N}$ is defined by:

$$e_t(j) = \begin{cases} S \text{ if } \delta_X(i,j) > t \\ I \text{ if } \delta_X(i,j) = t \\ R \text{ if } \delta_X(i,j) < t \end{cases}$$

Thus the contacts causing an infection (if none of the two nodes is not already recovered) are marked by the graph edges. The independance of the contacts leads to the independance of the Bernoulli variables and the uniformity of the probability to have an infectious contact between any two individuals, to the equality of the parameters of the Bernoulli variables, which hence gives the model $\mathscr{G}(n,p)$.

The epidemy spreads from the node $i$ to its direct neighbours, and iteratively to the upper layer of neighbours at each time unit. The connected component of $i$ in $X$ therefore represents the set of

[2] Proposed by Lowell Reed and Wade Hampton Frost in the twenties.


 ISSN: 2102-6238

individuals which will be infected over time. The study of the connected components of the model $\mathscr{G}(n,p)$ and of their size can then be interpreted from the point of view of the epidemiology and answers the question of the proportion of the population affected by the disease. The main lines of this study are presented in the next subsections.

We can so far notice that the assumptions made about the model dynamics are very simplistic and hardly plausible. Even though the Erdős-Rényi model has actually not the features of the networks observed in reality, it provides a very convenient framework, which allows a deep study. It is a necessary step before elaborating and understanding more complex models.

### 2.3. *Giant component in the model* $\mathscr{G}(n,p)$

In this paragraph, we are considering the Poissonian asymptotic framework, that is, the case where the average degree is a fixed parameter with respect to $n$; here we are considering the model sequence $\mathscr{G}(n,p_n)$, with average degree [3] $np_n = \lambda$. The name of this asymptotic framework comes in particular from the fact that the degrees then converges in distribution towards a Poisson distribution with parameter $\lambda$. The connected components sorted by size in descending order are denoted by $\mathscr{C}_{(1)},\mathscr{C}_{(2)}\dots$. The main result is the phase transition of the size of the largest component $\mathscr{C}_{(1)}$ with respect to the parameter $\lambda$. The following theorems tell how the largest component behaves (see in particular Draief and Massoulié, 2010 and other references cited afterwards):

**Theorem 2.1.**

**Subcritical regime** $\lambda < 1$**.** *There exists a constant* $A_1 \geq 0$ *depending on* $\lambda$ *such that:*

$$\lim_{n\to\infty} P_{\mathscr{G}(n,p_n)}\left(|\mathscr{C}_{(1)}| \leq A_1 \log(n)\right) = 1$$

**Supercritical regime** $\lambda > 1$**.** *Let* $p_{GW}(\lambda)$ *be the unique solution of* $x = e^{-\lambda(1-x)}$ *in* $]0,1[$*. There exists a constant* $A_2 > 0$ *depending on* $\lambda$ *such that for all* $\varepsilon > 0$*:*

$$\lim_{n\to\infty} P_{\mathscr{G}(n,p_n)}\left(\left|\frac{|\mathscr{C}_{(1)}|}{n} - (1-p_{GW}(\lambda))\right| \leq \varepsilon,\ |\mathscr{C}_{(2)}| \leq A_2\log(n)\right) = 1$$

**Critical regime** $\lambda = 1$**.** *There exists* $A_3 > 0$ *such that for all* $b > 0$*:*

$$P_{\mathscr{G}(n,p_n)}\left(|\mathscr{C}_{(1)}| \geq bn^{2/3}\right) \leq \frac{A_3}{b^2}$$

In the subcritical regime, the size of all components is at most logarithmic with respect to $n$. In the supercritical regime, it is the same, except for the largest component, the size of which is linear; there is thus an unique giant component a.a.s.. A non-negligible proportion of nodes is contained in this component, and this proportion is $1-p_{GW}(\lambda)$, where $p_{GW}(\lambda)$ is the extinction probability of a Galton-Watson process whose spring distribution is a Poisson distribution with parameter $\lambda$. The technical step of the proof using this process thus appears explicitly in the result.

[3] The average degree is rigorously $(n-1)p_n$ in this model, but both are asymptotically equivalent when $n$ tends to infinity.


 ISSN: 2102-6238

It is also visible in the formulation of the Reed-Frost model. A Galton-Watson tree is actually constructed through the graph: the root is the node where the epidemy had started from, and the children are the neighbors of this node in the graph (the infected nodes) and so on. The critical regime is more complicated and we give only this simple upper bound (Nachmias and Peres, 2010) to have an idea about the order of magnitude $n^{2/3}$ of the size of the largest component. More detailed results are presented in Janson et al. (1993); Bollobás (2001). Aldous (1997) also analyzes the connected components in this regime and establishes links between their sizes and the excursions of Brownian motions.

### 2.4. *Connectedness in the model* $\mathscr{G}(n,p)$

One of the first results about connectedness in the model $\mathscr{G}(n,p)$ is given in Gilbert (1959). Gilbert provides an exact and non-asymptotic formula of the probability of connectedness under this model, denoted by $\pi_{n,p}$. He uses an usual method in combinatorics, based on formal power series. He first writes $\pi_{n,p}$ as a function of $C_{n,l}$ ($l = n-1,\dots,N$), defined as the number of connected graphs with $n$ nodes and $l$ edges. Remind that if still $N = \binom{n}{2}$, each graph with $n$ nodes and $l$ edges is drawn with probability $p^l q^{N-l}$ in this model, where $q = 1-p$. Moreover a connected graph with $n$ nodes has at least $n-1$ edges (it is even a tree if it has exactly $n-1$ edges), therefore $C_{n,l} = 0$ for all $l < n-1$. Hence:

$$\pi_{n,p} = \sum_{l=n-1}^{N} C_{n,l} p^l q^{N-l}$$

Gilbert writes a formal power series for the double series $(C_{n,l})_{n\in\mathbb{N}^*,l\in\mathbb{N}}$ and converts it to a formal power series for $\pi_{n,p}$. By using Taylor series expansions and denoting $\mathscr{R}_n$ the set of the partitions [4] of the integer $n$, Gilbert obtains:

$$\pi_{n,p} = n! \sum_{(r_1,\dots,r_n)\in\mathscr{R}_n} \frac{(-1)^s (s-1)! q^{(n^2-1^2 r_1-\cdots-n^2 r_n)/2}}{r_1!\dots r_n!(1!)^{r_1}\dots(n!)^{r_n}}$$

where in the sum, $s = r_1+\cdots+r_n$. $\pi_{n,p}$ can be hence written as a sum whose number of terms is the cardinality of $\mathscr{P}_n$. However this number grows very fast with respect to $n$ and the formula cannot be computed even for small values of $n$.

#### *Connectedness and isolated nodes*

What should be remembered about connectedness in $\mathscr{G}(n,p)$ is that isolated nodes are its best enemy: when a graph under this model is not connected, it is mostly because of isolated nodes. Firstly, the probability of having one isolated node is obviously a lower bound of the probability of disconnectedness: indeed if there is an isolated node, it is actually a connected component for itself, and therefore the graph is not connected. Even though this bound can seem rough, these probabilities are actually equivalent when $n$ tends to infinity; if $I_n$ denotes the event "there is at least one isolated node":

[4] A partition of the integer $n$ is a $n$-tuple of integers $(r_1,\dots,r_n)$ such that $r_1+2r_2+\cdots+nr_n = n$


 ISSN: 2102-6238

**Proposition 2.1.**

$$P_{\mathcal{G}(n,p)}(I_n) \sim_{n\to+\infty} 1-\pi_{n,p} \sim_{n\to+\infty} nq^{n-1}.$$

The proposition is proved in Channarond (2013), using both upper and lower bounds of the probability of disconnectedness from Gilbert (1959).

As a summary, when $n$ tends to infinity, the disconnectedness essentially comes from isolated nodes and not from a lack of connections between two big components. A short explanation for this is that the probability that a given set of $k$ nodes is not connected to its complementary set is $q^{k(n-k)}$, where $k(n-k)$ is the number of possible edges between the set and the complementary one. And the function $x \mapsto x(n-x)$ grows fast in 1 and reaches its maximum in $n/2$, so that there are many more links to break to separate a component whose size is $k > 1$ than an isolated node. The same phenomenon may arise in models very different from Erdős-Rényi, for instance in random geometric graphs (see definition and comments in Subsection 4.2).

*Critical regime for the connectedness*

Gilbert inequalities allow to conclude to almost sure connectedness in the fixed average density regime $p > 0$. In order to find the critical regime for the connectedness between this regime and the regime $p = 0$ where the graph is empty, we consider that the average density vanishes when $n$ grows: $p = p_n \to 0$. The asymptotic framework is changed and the precision of the bounds is deteriorated. In particular the lower bound never allows to conclude to almost sure asymptotic disconnectedness. The upper bound allows nevertheless to bound the critical regime from below:

$$\text{If } \underline{\lim}\, \frac{np_n}{\log(n)} > \frac{4}{3}, \text{ then } P_{\mathcal{G}_{n,p_n}}(X \text{ connected}) \xrightarrow[n\to\infty]{} 1$$

To get the exact critical regime, isolated nodes are the key once again: we have to work more on them, and in particular on the limit distribution of their number $N_{\text{isol}}$, the expectation of which is:

$$\mathbb{E}\,(N_{\text{isol}}) = nq_n^{n-1} = \exp\left(-(np_n - \log(n)) + o(np_n)\right)$$

The crucial quantity is hence $c_n = np_n - \log(n)$, and we distinguish between two different regimes, according to whether the average degree $np_n$ dominates $\log(n)$ or not, i.e. $c_n \to +\infty$ or $c_n \to -\infty$. In the earlier, the number of isolated nodes shoots up, whereas in the latter it tends to zero.

**Limit distribution of the number of isolated nodes.** Let $c \in \mathbb{R}$ and $p_n = \frac{\log(n)+c}{n}$. Erdős and Rényi proved with the factorial moments method that in that asymptotic framework, $N_{\text{isol}}$ converges in distribution to a Poisson distributed variable with parameter $e^{-c}$, and therefore obtain the following proposition:

**Proposition 2.2.** *(Erdős-Rényi)*

$$P_{\mathcal{G}(n,p_n)}(N_{isol} = 0) \xrightarrow[n\to\infty]{} e^{-e^{-c}}$$


 ISSN: 2102-6238

This result has a more modern and simpler proof with the Stein-Chen's method, using two moments only (see Arratia et al., 1989 for the general method, and for example Draief and Massoulié, 2010 for the proof of the proposition 2.2 with this method). Introduced by Stein in 1972 so as to bound from above the total variation distance to a normal limit distribution, and then generalized by Chen in 1975 for a Poisson distributed limit distribution in Chen (1975), it is an usual tool in graphs, (see for example Penrose, 2003) and stochastic processes in a context of weak dependences, which is the case here in particular for the degrees. See also Barbour et al. (1992) for an overview, applications and references.

**Components larger than 2 nodes vanish.** Still in this asymptotic framework, one can show that the probability that the random graph under $\mathcal{G}(n,p_n)$ has a component larger than two nodes tends to zero. Note that it is sufficient to deal with the components whose size is less than $n/2$, indeed if there is no component smaller than $n/2$ nodes, there cannot be either any component larger than $n/2$, because the complementary node set would form components smaller than $n/2$. As a consequence, the disconnectedness still comes from isolated nodes in this asymptotic framework. Thus we can give the following theorem, called "double exponential":

**Proposition 2.3.**

$$\lim_{n\to\infty} P_{\mathcal{G}(n,p_n)}(G \textit{ connected}) = \lim_{n\to\infty} P_{\mathcal{G}(n,p_n)}(N_{isol}=0) = e^{-e^{-c}}$$

It implies that $\log(n)$ is the critical rate for the average degree $np_n$ for the connectedness:

**Théorème 2.1.** *(Erdős-Rényi) Let* $c_n = np_n - \log(n)$.
- *If* $c_n \to +\infty$, *then* $P_{\mathcal{G}(n,p_n)}(X \textit{ connected}) \xrightarrow[n\to\infty]{} 1$.
- *If* $c_n \to -\infty$, *then* $P_{\mathcal{G}(n,p_n)}(X \textit{ connected}) \xrightarrow[n\to\infty]{} 0$.

We can also say that, if the parameter $\mu = \frac{np_n}{\log(n)}$ is fixed, the connectedness has a phase transition with respect to the parameter $\mu$ in 1: if $\mu > 1$, the graph is a.a.s. connected, whereas if $\mu < 1$, it is a.a.s. disconnected. But the theorem is actually a bit more precise, since it clarifies what happens in some critical cases $\mu = 1$, in particular cases where $c_n$ tends to infinity not as fast as $\log(n)$ (*fast* hence remains relative).

Let us note that this phase transition arises at a faster density rate than that of the giant component. Basically more edges are required to connect all nodes instead of just a fixed proportion of them. But this extra effort is actually weak, because the rate for $p_n$ is hardly faster, $n^{-1}\log(n)$ instead of $n^{-1}$ in the Poissonian framework.

*Diameter*

Let us consider the regime such that $\frac{np_n}{\log(n)} \to +\infty$. As mentioned before, under $\mathcal{G}(n,p_n)$, the random graph is a.a.s. connected. In the case where the graph is connected, for any two nodes, there exists a path in the graph which connects these nodes. The question is what is the longest path, i.e. the largest distance between two nodes in the graph, also defined as the graph diameter. The following theorem adapted from Draief and Massoulié (2010) tells that in this regime the graph diameter under $\mathcal{G}(n,p_n)$ grows slower than logarithmically with respect to $n$, when it tends to infinity:


 ISSN: 2102-6238

**Theorem 2.2.** *It is assumed that* $\frac{np_n}{\log(n)} \to +\infty$. *There exists a constant* $B \geq 0$ *such that:*

$$\lim_{n\to\infty} P_{\mathscr{G}(n,p_n)}\left(\operatorname{diam}(X) \leq B\frac{\log(n)}{\log\log(n)}\right) = 1$$

### *2.5. Back to modeling*

The Reed-Frost model is based on two simplistic assumptions, which amounts to those of Erdős-Rényi model: independance of contacts, uniformity of the probability of infectious contacts between all individuals. The model can be interpreted in a similar fashion in social networks. Edges represent social interactions instead of infectious contacts. More precisely:

**Independance of interactions** Any individual $i$ connect to any other independently of all other connections, including those starting from $i$. However it does not seem plausible in a social network: in general friendships and meetings bring other friendships.

**Uniformity of interaction probability** The connection probability between any two nodes is the same, which does not seem either plausible. In particular, it implies that any individual can interact with any other with the same probability, despite their geographic distances or their sociological differences. Moreover in this model, each individual has the same average number of friends (degree), whereas people may have abilities to socialize.

In a global point of view, the average Erdős-Rényi random graph has no particular structure: at a fixed edge number, all topologies have the same probability to be drawn. This assumption is not consistent with the construction of social networks, where individuals have different social roles which shape the topology of the network. Besides in protein-protein biochemical interaction network, each protein plays a special role in the global regulation process, depending on its own molecular structure. Thus Erdős-Rényi model accounts neither for features of the individuals, likely having an effect on its interactions, nor the underlying dependence structure.

As a summary, this model is a reference, owing to all we know about it. It turns out to be a good technical tool, useful as elementary object in more elaborated models. As its main property is that its topology is fully random, it is used as a null hypothesis model in statistical tests as well, to detect remarkable features of graphs. The question is whether they may arise in a fully random topology, or they are significantly different. In such a test, it is crucial to know deeply the model, since we need to know how behaves the test statistic under the null hypothesis to construct the rejection region (see an example in Channarond, 2013 for a heterogeneity test and in Arias-Castro and Verzelen, 2013 for communities detection). Nevertheless the simplistic assumptions of the model are at odds with elementary observations, and the model turns out to be unsuitable to model real-world interaction networks.

## 3. Empirical features of real-world networks

After having shaken up the Erdős-Rényi model, we can wonder what should be the expected features from a random graph model, by describing empirically real-world networks. Review articles Albert and Barabási (2002) and Newman (2003) made a large list of empirical features of different interaction networks, especially depending on the application field. Despite their diversity, they can nevertheless share common features, like these following.


 ISSN: 2102-6238

**Small world** This feature has become famous since the experiment of the sociologist Stanley Milgram (Milgram, 1967), called "six degrees of separation". On the one hand it showed that in a friendship network, each person was not further than six handshakes from any other in the world. It seems to be much faster than expected, which explains why this concept is called "small world". In terms of graphs, it means that the smallest path connecting any two people was not longer than six nodes. A model is told to satisfy the small-world feature whenever the average distance between two nodes in the graph is dominated by $\log(n)$. Note that Erdős-Rényi model $\mathscr{G}(n,p_n)$ satisfies this feature when $\frac{np_n}{\log(n)} \to +\infty$, but by accident in a sense: its topology is fully random whereas people networks are spatially based. There are simply enough edges in this regime to produce many short paths.

On the other hand, Milgram's experiment also showed that finding the optimal path between two people does not require to know the whole network: an approximate idea of the target and the local view of the network each node has, are enough. This extra capacity offered by the network is called efficient routing.

**Density** Density of real-world networks generally vanishes when $n$ tends to infinity, which shows the limited capacity of the individuals to connect. In other words, the number of actual relationships grows strictly slower than the number of possible relationships. A random graph model is said to be sparse whenever $\zeta(X) = \mathscr{O}\left(\frac{1}{n}\right)$ almost surely or in expectation when $n$ tends to infinity.

**Connectedness** Real-world networks often have a large group of individuals interacting, and possibly some small separated groups. Therefore they are generally connected or have a giant component at least. Many random graph models are connected by construction. Note that Erdős-Rényi model $\mathscr{G}(n,p_n)$ can both be sparse and have a giant component a.a.s., but not be a.a.s. both sparse and connected: the connectedness costs too many edges so that sparsity remains.

**Degree distribution** Much attention is paid to this distribution, and so much that some models were created to display a prescribed distribution: these are called configuration models (Bender and Canfield, 1978; Molloy and Reed, 1995). Although many people wrote about degrees, hoping to find an universal distribution, it seems that it crucially depends on the type of the network. The discussion essentially focuses on the tail of the distribution, heavy or exponential. In the case of a heavy tail, the proposed distributions are called *scale-free*, and correspond to power laws. A random variable is said to follow a power law if its distribution $\mu$ is such that for all $k \in \mathbb{N}$:

$$\mu([k,+\infty[) = \alpha k^{-\beta}$$

where $\alpha, \beta \geq 0$ are parameters. Models whose degrees follow a power law are called scale-free by extension. To the best of our knowledge, scale-free models have been first mentioned in an article of Lotka in 1926 (Lotka, 1926) about researcher collaboration network. Their recent resurgence and the major enthusiasm they provoked come from the study of the World Wide Web, whose empirical degree distribution seems to have an heavy tail.

Even though power law is not universal as mentioned by Albert and Barabási (2002) in General Questions, it is the characteristic of a number of hubs in the network, i.e. nodes with a very high degree. In reality, some examples display a power law with an exponential cut, preventing too strong hubs.


 ISSN: 2102-6238

The degree distribution of Erdős-Rényi has an exponential tail. However Albert and Barabási (2002) remarks that even in the case of real-world networks with an exponential tail, the degree distribution is far from a Poisson like in the sparse Erdős-Rényi model.

**Transitivity** In a social network, this feature is illustrated by the common phrase "friends of my friends are my friends". If $i, j$ are two nodes, it means that the probability of connection of $i$ and $j$ is larger conditionally on the fact that they both are connected to a third distinct node $k$:

$$P(X_{ij} = 1 \mid X_{ik} = 1, X_{jk} = 1) > P(X_{ij} = 1).$$

One of the characteristics of a transitive model is the large number of triangles in the graph. This tendency is measured by the clustering coefficients, defined respectively in Watts and Strogatz (1998) and Bollobás and Riordan (2003) as follows:

$$CL_1(X) = \frac{1}{n} \sum_{i \in [n]} \zeta(X_{\sharp i}) \text{ and } CL_2(X) = \frac{\sum_{i \in [n]} \binom{D_i^X}{2} \zeta(X_{\sharp i})}{\sum_{i \in [n]} \binom{D_i^X}{2}} \tag{1}$$

where $X_{\sharp i}$ is the subgraph of $X$ induced by the neighbors of $i$ in $X$. $CL_2(X)$ is actually three times the number of triangles in the graph divided by the number of edge pairs sharing a common node. Both $CL_1(X)$ and $CL_2(X)$ are zero if $X$ is a tree and one if $X$ is a clique, but they are not equal in general. If $(X_n)$ is a graph sequence, the clustering coefficient $(CL_i(X_n))_n$, $i = 1, 2$ is said to be large if:

$$\liminf_{n \to \infty} CL_i(X_n) > 0 \text{ a.s.}$$

**Other features** There are many other expected features in the models depending on the application field, like the robustness of the connectedness against attacks deleting nodes in the graph in computer science (see Bollobás and Riordan, 2004a), the presence of some motifs of interest in biology (Birmele, 2012), etc.

## 4. Some heteregeneous models with generative processes

Many modelers, in particular physicists, have attempted to understand interaction mechanisms so as to get them into models. The goal is to reproduce global real-world local features and show that these mechanisms account for them. In this section, we present some fruitful attempts.

### *4.1. Small-world models of Watts-Strogatz and Kleinberg*

The Watts-Strogatz model (Watts and Strogatz, 1998) was designed to explain how small-world and large clustering coefficient can arise together in social networks. Indeed this combination cannot come from a random topology, since Erdős-Rényi model is small-world, but the clustering coefficient is not high. Moreover Kleinberg added an explanation of the efficient routing in Kleinberg (2000) after providing a rigorous algorithmic definition.

The common idea of the original model and the successive versions (to which Kleinberg's belongs) is to randomly perturb a regular graph, based on a lattice for example, by adding


 ISSN: 2102-6238

randomly edges, acting as shortcuts. In Watts and Strogatz (1998) shortcuts are created by redirecting randomly existing edges of the regular graph. In Newman and Watts (1999a,b), extra edges are added to the regular ones. Each node creates with probability $p \in ]0,1]$ a shortcut to a node, drawn uniformly among all other nodes (see Figure 1). These shortcuts reduce the diameter of the original regular graph, and is dominated by $\log(n)$ (see also Draief and Massoulié, 2010). Newman and Watts (1999a,b) studies the phase transition [5] between the linear regime for the average distance in the graph and the emergence of the small-world property.

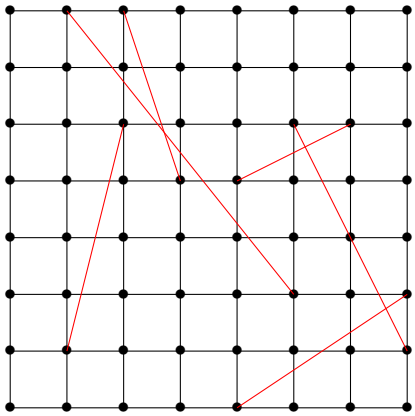

FIGURE 1. *Realization of a random graph of the Watts-Strogatz model on the grid* $7 \times 7$. *From each node a shortcut is created with probability* $p = 0.07$, *to a node uniformly drawn. Grid edges are coloured with black and shortcuts with red.*

However these models do not explain the efficient routing of the Milgram's experiment. Kleinberg (2000) proved that a decentralized algorithm, i.e. which knows the regular graph and the edges connected to the nodes it goes through only, cannot find paths shorter than a power of $n$ in average, which is much longer than the average distance (here logarithmic with respect to $n$).

Kleinberg's model is the following. Nodes correspond to the points of the grid $[m] \times [m] \subset \mathbb{R}^2$, there is hence $n = m^2$ nodes. Nodes closer than $D \in \mathbb{N}$ according to the $L^1$-distance, are connected in a deterministic manner. Thus far the graph diameter is $\mathcal{O}(n/D)$. Random shortcuts are then added: $r \in \mathbb{N}$ edges start from each node. For all nodes $i \neq j$, the probability that the end of an edge starting from $i$ is $j$, is inversely proportional to some power $\alpha \geq 0$ of their $L^1$-distance:

$$P(E = j) = \frac{\|i-j\|_1^{-\alpha}}{\sum_{k \neq i} \|i-k\|_1^{-\alpha}} \text{ where } \alpha \geq 0.$$

If $\alpha = 0$, the edges are uniformly drawn and we get a model C la Watts-Strogatz. Kleinberg (2000) proves that for the parameter $\alpha = 2$ and only for this value, the greedy [6] algorithm finds paths whose average length is dominated by $\log^2(n)$. The model with parameter $\alpha = 2$ is hence the only one navigable.

[5] The scheme where extra edges are added (instead of redirected) make this study easier from a technical point of view.

[6] The term greedy is the general name of an iterative optimization strategy, consisting in optimizing locally at each iteration, hoping to reach the global optimization at the end of the process.


 ISSN: 2102-6238

These models have a transparent interpretation from the point of view of social networks. For example in the Kleinberg model, the grid provides a geographic base for the network, and the regular graph models the neighborhood relationships. The penalization of the probability of a shortcut by the distance of its ends is also explained by the fact that geographic distance limits the number of relationships. The parameter $\alpha$ can be interpreted as the inertia level of the individuals. The higher it is, the less far the shortcuts go. The smaller it is, the further they can go, up to the critical value $\alpha = 0$, where they can go anywhere with the same probability.

The first clustering coefficient of small-world models is high, and the degree distribution is not heavy-tailed according to Newman and Watts (1999a). Basing the graph on a regular graph is two-edged: on the one hand, the spatial structure brings heterogeneity but on the other hand it severely limits the variety of the topologies of the final graph. This regularity, as well as the uniformity of the shortcut distribution, does not take into account individual social characteristics.

### *4.2. Random geometric graphs*

In the class of random geometric graphs, random positions $(Z_i)_{i\in[n]}$ taking values in a space equipped with a distance $\delta$ are associated with the nodes. Edges are created between nodes closer than some threshold $r > 0$. They are thus deterministic conditionally on the positions; for all $i \neq j$:

$$X_{ij} = \mathbb{1}_{\delta(Z_i,Z_j)\leq r}$$

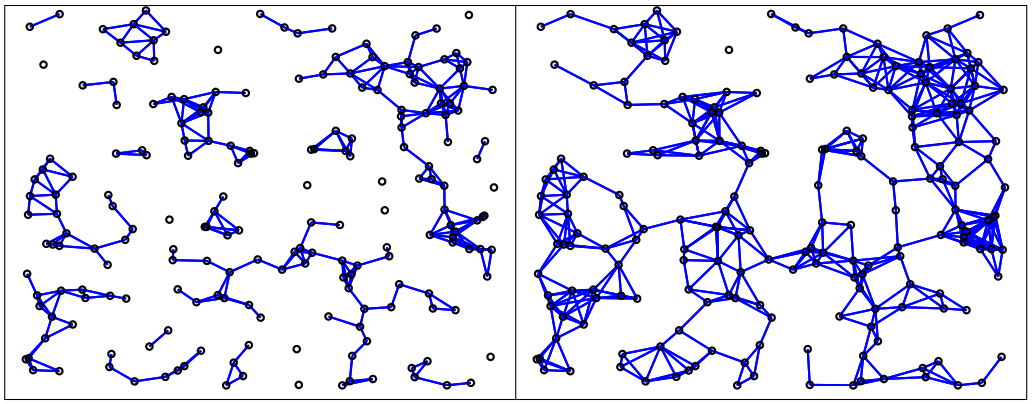

FIGURE 2. *Realization of two random geometric graphs with* $n = 200$ *on the unit square of* $\mathbb{R}^2$ *with an uniform density,* $r = 0.07$ *(left) and* $r = 0.1$ *(right). Positions* $(Z_i)_{1\leq i\leq n}$ *are the same in both graphs, only the threshold r changes.*

This class of models is useful for modeling propagations of a fire in a forest or of a disease, or wireless networks for instance (Haenggi et al., 2009). These graphs are largely studied in Penrose (2003), especially in the case where the positions are drawn with an homogeneous Poisson process on $\mathbb{R}^d$ or with an uniform density on $[0,1]^d$. Let us note the result of Gupta and Kumar (1998), proving that the probability of connectedness of the uniform random geometric graph on the unit ball of $\mathbb{R}^2$ is equivalent to the probability of having an isolated node when $n$ tends to infinity, exactly like in Erdős-Rényi model. Moreover the critical regime on the average degree for the connectedness is the same as in Erdős-Rényi model, despite the different nature of these models.


 ISSN: 2102-6238

A proof made in Channarond (2013) makes a link between the two classes of models: locally (in a geometric sense), random geometric graphs contain an Erdős-Rényi random graph.

### *4.3. Barabási-Albert model*

Almost at the same time as Watts and Strogatz at the end of the nineties, Albert and Barabási introduced a scale-free model in Barabási and Albert (1999). One of the characteristics of their model, and of all models in the class of growth models, is that they are iteratively constructed, adding at each step a new node to those already present. From a physical point of view, the growth dynamics of the graph is described at a microscopic scale, highlighting the topological mechanisms of the networks, leading to their macroscopic features.

In this model the specific mechanism is called preferential attachment; it consists in linking the new node preferably to popular nodes (those with a high degree), which thus explains how hubs and the power law emerge. Introduced in 1925 in Yule (1925) to model researcher citation networks, this iterative scheme came recently back into the spotlight to model the evolution of the Web. The model of Barabási and Albert (1999) is defined in a rigorous manner and called Linearized Chord Diagram (LCD) in Bollobás et al. (2001).

In the following version of Draief and Massoulié (2010), which depends on a parameter $p$, the degree distribution converges to a power-law distribution with parameter $(3-p)/(1-p)$, covering values from 3 to $+\infty$. We are getting started from a base graph $X^0 = (V_0, E_0)$. Then we are constructing by induction the sequence $(X^n)_{n\in\mathbb{N}}$, where for all $n \in \mathbb{N}$, $X^n = (V_n, E_n)$. At the step $n \in \mathbb{N}^*$, a node $i \notin V_{n-1}$ is added, from which an edge starts. Therefore $V_n = V_{n-1} \cup \{i\}$, the number of nodes at this step is $n + |V_0|$ and the number of edges $n + |E_0|$. The other end of the edge is uniformly drawn in $V_{n-1}$ with probability $p \in [0,1[$ and is drawn with probability $1-p$ according to the preferential attachement rule, defined as follows; for all $j \in V_{n-1}$, the probability of connection between $i$ and $j$ is:

$$P(X_{ij}^n \mid X^{n-1}) = \frac{D_j^{X^{n-1}}}{2(|E_0|+n)}$$

The degree distribution is scale-free and the model satisfies the small-world property, thanks to the hubs acting as routers, through which many shortest paths can go. However its second clustering coefficient is small.

The tail of the degree distribution is even heavier when $p$ is higher, insofar as the exponent of the power-law distribution decreases. By construction, this parameter controls the proportion of preferential attachment introduced at each step, and then the intensity of hub creation. The LCD [7] model amounts to $p = 0$. Bollobás and Riordan (2003) established a review of the results about the original LCD model.

Let us remark that heterogeneity is brought into the graph by an unbalanced mechanism which automatically forces the degrees of some nodes to deviate. This heterogeneity does not depend

[7] Note that in Bollobás and Riordan (2004b); Bollobás et al. (2001); Bollobás and Riordan (2003), when a new edge is created, the node $i$ itself is also considered, which may create self-loops. This enables to describe the distribution of the random graph in an easier manner. Moreover, these articles consider the creation of more than one edge from the new node as well.


 ISSN: 2102-6238

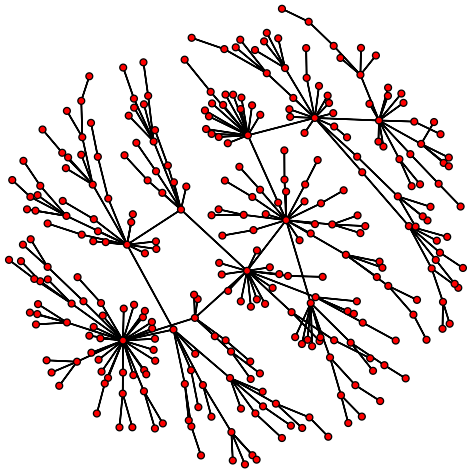

FIGURE 3. *Realization of a random graph from Albert and Bárabasi's model, with* $n = 300$*, constructed from a graph* $X_0$ *with 10 nodes, which is itself a realization of Erdős-Rényi graph with* $p = 0.3$*.*

on a parameter determined before and based on the node characteristics. It is rather the opposite: any new node actually follows the same story: the distribution of their edge depends only on the present graph, but not on any individual parameter.

A generalization of this model is possible by using other attachment functions which are increasing with respect to the degree, but other than linear, see Dommers et al. (2010) and Buckley and Osthus (2004).

Models with copying processes are another usual subclass of growth models. At each step, an existing node is uniformly drawn and the added node "copies" it by connecting to (some of) its neighbours. Kleinberg et al. (1999) addresses this kind of mechanisms, originally inspired by the way how Web pages are created. It had a great success at modeling the genome evolution as well (see Chung et al., 2003 for example).

### *4.4. Conclusion*

These models sketch the mechanisms of the formation of local topological structures, to explain in the simplest way how the global features of the real-world networks emerge. They are generally the fruit of an *ad hoc* approach, which successfully answers questions in physics in particular.

But despite the obvious heterogeneity of these models, all nodes share a common construction process, which forgets possible individual characteristics having effects on interactions. It could be told that all individual parameters are like replaced by a few parameters averaged over the whole population. Variability of these models is thus limited and their topology is rather rigid. The book of Kolaczyk (2009) mentions that it does not prevent from estimating these parameters, often having an interesting interpretation in the applications. However these models generally turn out to be not very suitable for a statistical study of variability, aiming at separating individual and global effects and then estimating them. Even though statistics needs models fitting well to observed data, these moreover must allow the reverse way of physics: going back from the global


 ISSN: 2102-6238

TABLE 1. *Regression models*

| | |
|---|---|
| Linear regression of the size on age | $X_i \sim \mathcal{N}(a+bz_i, \sigma^2)$ |
| observed covariate | $z_i$ age of the individual $i$ |
| parameters | $a$ average size at birth, $b$ growth rate by time unit |
| Logistic regression of the interaction on distances and sociability | $X_{ij} \sim \mathcal{B}\left(\frac{e^{\alpha+\beta(\eta_i+\eta_j)-\gamma d_{ij}}}{1+e^{\alpha+\beta(\eta_i+\eta_j)-\gamma d_{ij}}}\right)$ |
| observed covariates | $\eta_i, \eta_j$ sociabilities of the individuals $i$ and $j$<br>$d_{ij}$ distance between $i$ and $j$ |
| parameters | $\alpha_0$ fixed effect, $\alpha_1$ weight attributed to the sociability<br>$\alpha_2$ weight attributed to the distance |

network to the individual characteristics of each node.

## 5. Regression models

### 5.1. Modeling heterogeneity in statistics

One of the main features expected from a statistical model is the goodness of fit to the observed data. For example if we model the adult size of the individuals of an asexual animal species, a normal distribution is convenient: it models efficiently a variable centered on the characteristic mean of the species, on which symmetric fluctuations — owing to the individual genetic and environmental variability — are superimposed. However goodness of fit is deteriorated if this model is used on all individuals of the species, regardless of the development stage, or if the species is sexual and the size of male and female are significantly different. The population is therefore not homogeneous anymore and the variability is not explained by fluctuations only, but by individual characteristics the model should include as well. In these simple examples, age or sex are generally available covariates and a linear regression of the size with respect to the age or an ANOVA with respect to the sex can be used as a first analysis.

### 5.2. Back to random graph models

In the field of networks, the Erdős-Rényi model corresponds to an homogeneous population. Gilbert used the model $\mathcal{G}(n,p)$ to model a phone network. But these networks are strongly affected by individual characteristics, like the so-called "telephone sociability", and geography. A regression can be implemented to take these observed covariates into account, the logistic framework being the most convenient for the case of binary graphs, involving Bernoulli distributed variables.

Let $(\eta_i)_{i\in[n]}$ be real numbers quantifying the sociability, negative values corresponding to people allergic to the phone, and positive values to addicted people. Here they are assumed to be observed covariates, based on the number and the duration of the calls for instance. Let $(d_{ij})_{i,j\in[n]}$ be the matrix of the distances between people, so that call probability is penalized by the geographic distance, like in Kleinberg model (see Subsection 4.1). Let $\alpha_0 \geq 0$ an unknown fixed effect, and $\alpha_1, \alpha_2 \geq 0$ unknown weighting parameters attributed respectively to the sociability and


 ISSN: 2102-6238

the distance. Random variables $X = (X_{ij})_{i,j\in[n]}$ are assumed to be independent, and the random graph model is defined as follows:

$$\text{logodd}(X_{ij} = 1) = \alpha_0 + \alpha_1(\eta_i + \eta_j) - \alpha_2 d_{ij} \tag{2}$$

or equivalently:

$$X_{ij} \sim \mathscr{B}\left(\frac{e^{\alpha_0+\alpha_1(\eta_i+\eta_j)-\alpha_2 d_{ij}}}{1+e^{\alpha_0+\alpha_1(\eta_i+\eta_j)-\alpha_2 d_{ij}}}\right)$$

Thus if $\eta_i$ is fixed, the larger $\eta_j$ is, the more likely the connection between $i$ and $j$ is. Moreover if $\eta_i$ and $\eta_j$ are fixed, the smaller their distance $d_{ij}$ is, the more likely the connection is.

$\alpha$ denotes the vector of the parameters $(\alpha_0, \alpha_1, \alpha_2)$. The likelihood of $X$ is for all graph $x$ with $n$ nodes:

$$P_\alpha(X = x) = \frac{1}{\Lambda_\alpha}\exp\left(\sum_{1\le i<j\le n} x_{ij}(\alpha_0 + \alpha_1(\eta_i+\eta_j) - \alpha_2 d_{ij})\right)$$

where $\Lambda_\alpha$ is the normalization constant. The estimation of the parameters $\alpha$ is simple and can be made by the maximum likehood method.

Note that this model amounts to an undirected version of the $p_1$ model introduced in Holland and Leinhardt (1981). It is identifiable under some noncolinearity assumptions, since the covariates $\eta_i$ and $d_{ij}$ are supposed to be known. Moreover if $\eta_1 = \eta_2 = \cdots = \eta_n$ and $d_{12} = d_{13} = \cdots = d_{n-1,n}$, the population is homogeneous again, and everyone lives at the same distance from each other. It actually amounts to the Erdős-Rényi model $\mathscr{G}\left(n, (1+e^{-(\alpha_0+2\alpha_1\eta_1-\alpha_2 d_{12})})^{-1}\right)$.

Regression models take thus account of heterogeneity, establishing an explicit relation between individual characteristics and the distribution of the observations (see Table 1). However a disappointing aspect is still the independance of the edges. Even though connection probabilities are related, for example via the triangle inequality in the case of the distances, the realizations of the edges themselves are not dependent in the usual probabilistic sense. Moreover, all relevant covariates accounting for heterogeneity have to be observed to establish such regressions, which is however not possible in general.

## 6. Exponential random graph models (ERGM): Dependences "C la carte"

The previous model (2) belongs to the more general class of the exponential random graph models (ERGM), also called $p^*$ models. These are random graph models, which can be written as models of the exponential family. In this subsection, we call configuration any set of edges, that is, a subset of $\mathscr{P}_n$, the set of pairs of $[n]$. ERGM models are those whose distribution can be written as follows:

$$P_\theta(X = x) = \frac{1}{\Lambda_\theta}\exp\left(\sum_{C\subset\mathscr{P}_n}\theta_C g_C(x)\right) \tag{3}$$

where $x = (x_{ij})_{i,j\in[n]}$ is an adjacency matrix associated to the graph $(V_x, E_x)$, and:

- $g_C(x) = \prod_{\{i,j\}\in C} x_{ij}$ (or equivalently, $g_C(x) = \mathbb{1}_{C\subset E_x}$), telling whether the configuration $C$ is present in $x$, i.e. whether edges it contains are present in $x$.


 ISSN: 2102-6238

– $\theta_C$ the coefficient associated with configuration $C$. If $\theta_C = 0$, edges of the configuration are mutually independent conditionally on edges of $\mathscr{P}_n \backslash C$, then this configuration will occur only by chance. On the contrary, if $\theta_C > 0$ (respectively $\theta_C < 0$) the configuration is boosted (resp. penalized) and will occur more often (resp. less often) than by chance: it brings some dependence between the edges of the configuration. Each coefficient $\theta_C$ can depend itself on one or more parameters, and also covariates, like in the example of the previous subsection.
– $\Lambda_\theta$ is the normalization constant of the distribution.

ERGM models thus allow to parametrize directly the tendencies of the random graph to produce some configurations. This is a real switch in the core of the network topology and its dependence structure: by allocating a non-zero coefficient $\theta_C$ to configurations $C$ with several edges, some dependence is created between them. Note that any vector $(\theta_C)_{C \in \mathscr{P}_n} \in \mathbb{R}^{\binom{n}{2}}$ does not necessarily provide any valid model: coefficients of configurations sharing edges are related and impose constraints on each other. For example, for $k \in [n-1]$, a configuration of $k$ edges sharing the same node is called $k$-star; if $C$ and $C'$ are respectively a $k$-star and a $k+1$-star such that $C \subset C'$, it is not possible to have both $\theta_C = 0$ and $\theta_{C'} > 0$. The presence of $C'$ in a graph implies that of $C$, i.e. the realization of a $(k+1)$-star requires that of a $k$-star. The Hammersley-Clifford theorem (Besag, 1974) clarifies the validity assumptions of such a model.

When non-zero coefficients are allocated to configurations larger and larger, the dependence structure is getting more and more complicated. The dependence graph $D = (V_D, E_D)$ is often used to display this structure: $V_D$ is the set of all random variables of the model, here $V_D = \{X_{ij};\ 1 \leq i < j \leq n\}$, identified to $\mathscr{P}_n$; then two nodes are connected in $D$ if the corresponding variables are dependent conditionally on all others. Note that for all configurations $C \subset \mathscr{P}_n$, $\theta_C$ does not equal zero if and only if $C$ is a clique in the dependence graph $D$.

In the model of the previous subsection, the only specified configurations were the singletons $\{\{i,j\}\}$, for all $i \neq j$, so that edges are all mutually independent. Equivalently, $D$ was the empty graph. Coefficient $\theta_{ij}$ associated to $\{i,j\}$ was $\theta_{ij} = \alpha_0 + \alpha_1(\eta_i + \eta_j) - \alpha_2 d_{ij}$, including covariates $\eta_i$, $\eta_j$ and $d_{ij}$, and parameter $\alpha$.

### *6.1. Markov random graphs*

Frank and Strauss (1986) proposed a general concept of Markovian dependence in the networks: for all $k \in [n-1]$, $k$ edges are dependent if and only if they share the same node, i.e. if they form a $k$-star. Frank and Strauss (1986) shows that models satisfying this paradigm are a subclass of ERGM, whose non-zero coefficients are those associated with configurations $C$ which are $k$-stars or triangles. They are called Markov graphs and in particular produce transitive models.

If distinct coefficients are allocated to the configurations, some heterogeneity is brought to the model: each configuration is given its own probability occuring. However, it is not feasible to allocate distinct coefficients to all possible configurations, since the model would be overparametrized and not identifiable. A reasonable assumption is the uniformity within isomorphic configuration classes, that is, allocating the same parameter to configurations which are isomorphic.

Configuration classes up to isomorphism are called motifs. As we have already done before, configurations can be thus classified depending on the class they are member of. Configurations containing only one pair of nodes are called edges, that formed by any three edges and involving


 ISSN: 2102-6238

exactly three nodes are called triangles, that formed of $k$ edges sharing one common node are called $k$-stars, etc.

In the case of Markov models, Frank and Strauss (1986) gives the same coefficient $\theta_1$ to all edges, $\theta_k$ to all $k$-stars for all $k \in \{2,\dots,n-1\}$, and $\theta_\tau$ to all triangles. The distribution of the resulting random graph model is hence, for all graph $x$ with $n$ nodes:

$$P_\theta(X=x) = \frac{1}{\Lambda_\theta} \exp\left(\theta_1 L(x) + \sum_{k=2}^{n-1} \theta_k S_k(x) + \theta_\tau T(x)\right)$$

with $L(x)$ the number of edges of $x$, for all $k \in \{2,\dots,n-1\}$, $S_k(x)$ the number of $k$-stars in $x$ and $T(x)$ the number of triangles. To prevent once again from having too many coefficients, some constraints can be added on $(\theta_k)_{2\le k\le n-1}$, for example either by assuming that the last ones are zero, or to reach finer effects, by assuming that they satisfy some given progression with respect to $k$ and depending on one parameter only, like proposed by Snijders et al. (2006).

Thus parameters $\theta_1,\dots,\theta_{n-1},\theta_\tau$ allow to set the appearance of motifs such as triangles and stars in the network. Snijders et al. (2006) also proposes to add extra dependences, i.e. more elaborated motifs. Instead of just neighboring edges, dependences between $X_{ij}$ and $X_{kl}$ where nodes $i,j,k,l$ are distinct could be taken into account when an extra edge connects an end of $\{i,j\}$ to an end of $\{k,l\}$, for instance if the edge $\{i,k\}$ is present.

As a conclusion, the class of ERGM essentially provides a good framework of heterogeneous random graphs for the statistician interested in motifs. Graph motifs can have some interpretation in applications; for example, triangle is the basic motif of a social network, and correspond to the principle "friends of my friends are my friends". A large number of triangles indicates that well-connected communities emerge within the population. This is an indicator on the mixing of the population, that is, how likely are the connections between friends of friends. In biology, for instance in protein-protein networks, some motifs correspond to regulation loops, and explain how the regulation processes work (Milo et al., 2002; Birmele, 2012). More generally, motifs can be thought of as universal bricks of more elaborated networks (Milo et al., 2002).

### *6.2. Group characteristics*

Nevertheless allocating a coefficient to each motif rather than to each configuration harms the heterogeneity of the network, since each coefficient depends on the motif, and not on the individuals involved in the configuration themselves. Some heterogeneity can be brought back without overparametrization by assuming the population to be split into groups and that coefficients depend on the motif and on the groups involved in the configuration. For example, considering the previous example of the phone network, an additional effect arises in the international scale: that of the borders, breaking the geographic regularity. Let us consider the case of two countries with the following model. The population is split into two parts: $[n] = G_1 \sqcup G_2$. Let $\gamma_1,\gamma_2 \ge 0$ the additive effect favouring national calls and $\gamma_{12} = \gamma_{21} \ge 0$ the soustractive effect penalizing


 ISSN: 2102-6238

international calls [8]. If $i \in G_k$ and $j \in G_l$, let us note:

$$\theta_{ij}^{kl} = \begin{cases} \alpha_0 + \gamma_1 + \alpha_1(\eta_i + \eta_j) - \alpha_2 d_{ij} \text{ if } k = l = 1 \\ \alpha_0 + \gamma_2 + \alpha_1(\eta_i + \eta_j) - \alpha_2 d_{ij} \text{ if } k = l = 2 \\ \alpha_0 - \gamma_{12} + \alpha_1(\eta_i + \eta_j) - \alpha_2 d_{ij} \text{ if } k \neq l \end{cases}$$

Thus the distribution of the random graph is as follows, for all graph $x$ with $n$ nodes:

$$P_\theta(X = x) = \frac{1}{\Lambda_\theta} \exp\left( \sum_{\{i,j\} \in \mathscr{P}(G_1)} \theta_{ij}^{11} x_{ij} + \sum_{\{i,j\} \in \mathscr{P}(G_2)} \theta_{ij}^{22} x_{ij} + \sum_{(i,j) \in G_1 \times G_2} \theta_{ij}^{12} x_{ij} \right)$$

We emphasize that the groups are supposed to be known, and are not unobserved variables. As well as regression models of Subsection 5.1, this setting is hence not designed for clustering, and more generally it does not allow to characterize the heterogeneous structure of the population. Other models where the groups are unobserved variables will be considered in Section 7.

### 6.3. *Inference under ERGM*

The exponential family often turns out to be a pleasant framework for inference; it is also true for models with latent variables, often estimated with the EM algorithm. In the case of independent identically distributed samples, the estimators based on the maximum likelihood have an explicit expression. The log-likelihood of an ERGM can be written as follows for all graph $x$:

$$\mathscr{L}(x, \theta) = \theta^T g(x) - \psi(\theta)$$

where $g$ is the vector containing all functions $g_H$, $\theta$ the vector containing all parameters $\theta_H$ for all possible configurations $H$, and $\psi(\theta)$ is the logarithm of the normalization constant $\Lambda_\theta$.

Let us remind that independent ERGM are those such that the only configurations specified are edges. In these, the framework amounts to that of the generalized linear model and inference is not problematic. On the other hand, and the issue will be the same with the EM algorithm, adding more dependences make the inference much more complicated, even in the basic case of Markov models. The method of the maximum likelihood is not feasible, because it requires to minimize the normalization constant $\Lambda_\theta$, which is a sum of $2^{\binom{n}{2}}$ terms.

An overview of the ERGM and aspects of the inference can be found in Kolaczyk (2009); Robins et al. (2007). Here we briefly explain the main inference strategies in the ERGM. Most theoretically advanced and most currently used strategies have been developed under many versions, essentially gathered in two main classes of methods, both based on MCMC. One of these, detailed in Hunter and Handcock (2006), consists in maximizing an approximation of $\mathscr{L}(x, \theta) - \mathscr{L}(x, \theta^{(0)})$, which is the logarithm of the likelihood ratio between any parameter $\theta$ and the fixed parameter $\theta_0$. It can be written down as follows, using algebraic properties of the exponential family:

$$\begin{aligned} \mathscr{L}(x, \theta) - \mathscr{L}(x, \theta^{(0)}) &= (\theta - \theta^{(0)})^T g(x) - (\psi(\theta) - \psi(\theta^{(0)})) \\ &= (\theta - \theta^{(0)})^T g(x) - \log\left( \mathbb{E}_{\theta^{(0)}} \left( \left( \theta - \theta^{(0)} \right)^T g(X) \right) \right) \end{aligned}$$

[8] The model is not identifiable, additional constraints on the effects are needed to estimate the parameters.


 ISSN: 2102-6238

The approximation is obtained by estimating the expectation above by MCMC. Another way to proceed (Snijders, 2002) consists in using Robins-Monro algorithm, described by Kolaczyk (2009) as a stochastic version of the Newton-Raphson algorithm.

The second class of methods is based on a pseudo-likelihood, and uses a property of the exponential model to turn the inference of the ERGM into the inference of a logistic model. It had been originally introduced in Besag (1975) and was adapted to ERGM in Strauss and Ikeda (1990). The set of the variables $X = (X_{ij})_{i,j\in[n]}$ except $X_{kl}$ (and $X_{lk}$) is denoted by $X^{\setminus\{k,l\}}$. Then we can write:

$$\text{logodd}_\theta\left(X_{kl} = 1 \mid X^{\setminus\{k,l\}} = x^{\setminus\{k,l\}}\right) = \sum_{\substack{C\subset\mathscr{P}_n \\ \{i,j\}\in C}} \theta_C \delta_C^{ij}(x) \tag{4}$$

where $\delta_C^{ij}(x)$ is the difference between the values of $g_C(x)$ when $x_{ij} = 1$ and $x_{ij} = 0$, as if the distribution of $(X_{ij})_{i,j\in[n]}$ was that of a logistic regression model on the $\delta_C^{ij}$, forgetting that each one is conditioned on $X^{\setminus\{i,j\}}$. Then the inference is processed as if it was really such a model, by forgetting the dependences between the $X_{ij}$. The following log-pseudo-likelihood is finally maximized:

$$\sum_{\{i,j\}\in\mathscr{P}_n} \log P_\theta\left(X_{ij} = 1 \mid X^{\setminus\{i,j\}} = x^{\setminus\{i,j\}}\right).$$

Wasserman and Robins (2005) explains how to compute the difference statistics $\delta_C^{ij}$, which is not computationally complex. Robins et al. (2007) recall that (4) is not a logistic model, as the $X_{ij}$ are not independent, which leads to possible bias and underestimated variance. Moreover this method is still theoretically limited as well, since the behaviour of the pseudo-likelihood is not fully understood (Besag, 2001; Snijders, 2002) yet. Once again, let us note the parallel with the EM algorithm, addressed later in Subsection 7.5. Approximating the likelihood by ignoring dependencies has also led to the approximate inference method called Variational EM. This approach has turned out to be fruitful in models with a very intricate dependence structure, whenever the limiting structure has no dependences anymore.

## 7. Statistical models with latent variables

### 7.1. *Modeling hidden data in statistics*

For many complex phenomena, some relevant variables accounting for the heterogeneity of the observations are not available, because they may be physically unattainable or maybe too costly to acquire. Extra random variables can be nevertheless considered and added to the model despite that they are not observed. Such variables are said to be *hidden* or *latent*. Therefore three different types of variables are to be distinguished in these models. Firstly, the *observed variables* $X$ are the data to analyze; in this article, $X = (X_{ij})_{i,j\in[n]}$ is a graph. Secondly, the *covariates* are also observed variables, but are deterministic and supposed to account for the heterogeneity and variability of the observations $X$. In our setting, they can describe individual effects of the nodes (in the model (2), these are the phone sociabilities $(\eta_i)_{i\in[n]}$) or pair effects (the distances $(d_{ij})_{i,j\in[n]}$ in model (2)). Finally, *hidden* or *latent variables* also account for the heterogeneity of $X$, but are random variables which are not observed. Therefore they cannot be directly used to infer the


 ISSN: 2102-6238

model unlike covariates. In this article, each node $i \in [n]$ is allocated a latent variable denoted by $Z_i$.

**Example.** The phone network in Belgium is characterized by two large and very connected communities (Blondel et al., 2010), which remain unexplained by the previous phone netwok model (2). Indeed, observed covariates like individual sociability and geographic distances are not sufficient to explain the splitting of Belgium. To model this heterogeneous structure, each node $i$ is allocated a latent random variable $Z_i$, being either equal to 1 if interlocutor is from the first community or equal to -1 from the second. Let $\gamma \geq 0$ be the weight of the effect of the structure. The model then becomes:

$$\text{logodd}(X_{ij} = 1) = \alpha_0 + \alpha_1(\eta_i + \eta_j) - \alpha_2 d_{ij} + \gamma Z_i Z_j$$

The product $Z_i Z_j$ increases or decreases the connection probability between nodes $i$ and $j$, according whether it equals 1 or -1. *A posteriori*, it is noticed that this latent variable actually corresponds to the main language of the nodes (french or flemish). The structure of the phone network of Belgium thus reflects the linguistic heterogeneity of Belgium.

*Statistical challenges*

Models with latent variables are designed to capture the underlying heterogeneous structure of the population via the latent variables, which are the own characteristics of each individual. Beyond the goodness of fit, the main motivation using such models is clustering, that is, getting back to the structure of the population from the observed graph $X$ only. The mathematical procedure generally consists in splitting the population into groups (clusters), such that any two individuals from the same group are as similar as possible, and any two from different groups are as dissimilar as possible. However, clustering is a vague notion, covering more than one idea: in particular, the similarity function is not universally defined.

Models with latent variables bring new challenging statistical questions. The first one is how to define the right notion of clustering matching with the data (see also the discussion in Paragraph 7.7), and then choosing a model matching with this notion. Models with latent variables reduce questions of clustering to inference of the model: once the modeling framework is given, the last but not the least thing to do is inference. Only from the observation $X$, is it possible to construct consistent methods in order to:

- test the existence of clusters or equivalently test whether the observation $X$ is significantly heterogeneous, and therefore whether it is relevant to add latent variables in the model *(hypothesis testing)*,
- find the best number of clusters fitting to the data *(model selection)*,
- split the population into "homogeneous" groups *(classification, or clustering)*,
- in the parametric case, estimate the distribution of the latent variables *(estimation)*.

Typically, these questions have been already answered for finite gaussian mixture models, and more generally models from the exponential family for individual data. Consistent methods have been already developed as well. EM algorithm (and numerous variants) is the most popular procedure to estimate models with latent variables in the parametric case. But this method is not used at all for random graph models, because it needs the conditional distribution of the latent


 ISSN: 2102-6238

variables $Z$ on the observed ones $X$ to be tractable, which is generally not satisfied by such models. This major issue is more discussed in Paragraph 7.5.

The question of testing heterogeneity should be first addressed when data are to be analyzed, but the topic is surprisingly rather absent in the literature, especially in network models. This topic is nevertheless addressed for the Stochastic Blockmodel, see Channarond (2013).

### 7.2. *A general framework for random graph models with latent variables*

The general model presented here is inspired from Bollobás et al. (2007). It first proposes to allocate independent variables $Z = (Z_i)_{i\in[n]}$ to the nodes, representing their own individual characteristics; then edges $(X_{ij})_{i,j\in[n]}$ of the graph are independently drawn conditionally on $Z$, so that these characteristics bring heterogeneity into the interactions. Bollobás et al. (2007) introduced the model from a probabilistic point of view, aiming at studying phase transitions of the big component like in Erdős-Rényi model (see Section 2), but in heterogeneous graphs and in a general enough framework. From Bollobás et al. (2007), we are just borrowing the model and reinterpreting it from the statistical point of view of latent variables: variables $Z$ are assumed to be latent, whereas the graph $X$ is the observed data.

**Definition 7.1.** *Any couple $(\mathscr{S},\nu)$ such that $\mathscr{S}$ is a separable metric set and $\nu$ a probability measure over $\mathscr{S}$ is called latent space. Moreover any symmetric measurable map $\kappa : \mathscr{S}\times\mathscr{S} \to [0,1]$ is called connection function on $\mathscr{S}$.*

The latent space corresponds to the space of the individual characteristics of the nodes, equipped with their distribution $\nu$. Then the connection function $\kappa$ establishes the link between the characteristics and the edge distribution explicitly: if the characteristics of two nodes are $z$ and $z'$, the probability of these nodes being connected is $\kappa(z,z')$. The symmetry of $\kappa$ means that for all $z,z'\in\mathscr{S}$, $\kappa(z,z') = \kappa(z',z)$ and comes from the fact that the graphs are assumed to be undirected. Therefore such a triplet $(\mathscr{S},\nu,\kappa)$ defines a random graph model. More precisely:

**Definition 7.2.** *A random graph $X$ with $n$ nodes is said to be a random graph with latent space, if there exists:*

- *a latent space $(\mathscr{S},\nu)$ and a connection function on $\mathscr{S}$ denoted by $\kappa$,*
- *a sequence of independent and identically distributed variables $Z = (Z_i)_{i\in[n]}$ taking values in $\mathscr{S}$ and whose common distribution is $\nu$,*

*and if the distribution of $X$ satisfies:*

- *the mutual independance of the edges conditionally on $Z$; for all graph $x$ with $n$ nodes:*

$$P(X = x \mid Z) = \prod_{1\le i<j\le n} P(X_{ij} = x_{ij} \mid Z)$$

- *for all graph $x$ with $n$ nodes and for all $i\neq j\in[n]$, the distribution of $X_{ij}$ conditionally on $Z$ satisfies:*

$$P(X_{ij} = 1 \mid Z) = \kappa(Z_i,Z_j).$$

Note that this model may be not identifiable in some cases. This issue will be addressed in the examples. Three major examples of models, which are actually subclasses of this general family, are addressed in Subsections 7.4, 7.6 and 7.8.


 ISSN: 2102-6238

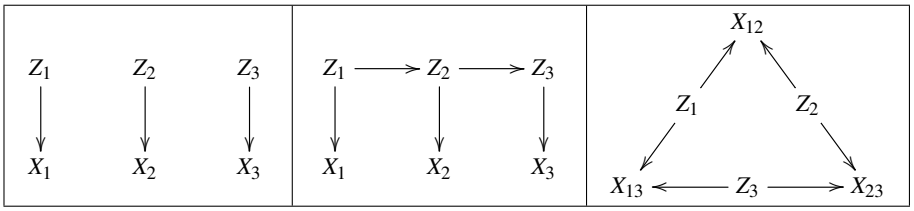


FIGURE 4. *Dependence graphs of models with latent variables: mixture models (left), Hidden Markov chain Models (middle), graph models (right).*

The family of models described by this definition seems to be reasonable from the point of view of modeling, but is also justified by a representation theorem of Lovász and Szegedy (2006), telling that all random graphs satisfying some natural assumptions are actually random graphs with latent variables (and even graphons, see Subsection 7.8).

### *7.3. Dependences in random graph models with latent variables*

*Dependences of the observed variables*

Let us compare models with latent variables in both situations: individual data versus interaction/graph data. $X$ denotes the observed variables and $Z$ the latent ones. In both cases latent variables $Z$ are independent, and observed variables $X$ are independent conditionally on $Z$. In spite of this, in the case of individual data observed variables $X$ are mutually independent (without conditioning) whereas they are not in the case of graph data. The difference lies in the structure of these data. In the setting of individual data $X = (X_i)_{i\in[n]}$, each observation $X_i$ involves only one individual $i$, and only one label $Z_i$, whereas the configuration of graph data $X = (X_{ij})_{1\le i,j\le n}$ is quite different: each observation $X_{ij}$ involves a pair of individuals $\{i, j\}$ (compare also dependence graphs in Figure 4). For example in a graph, any observations $X_{ij}$ and $X_{ik}$ sharing one common node are not independent, since they both depend on $Z_i$. $(X_{ij})_{i,j\in[n]}$ are thus not mutually independent, but any family of edges sharing no node is nevertheless mutually independent. The dependence structure is actually similar to that of Markov models (see Section 6): two edges are not independent if and only if they share one common node.

Therefore variables $(X_{ij})$ are not mutually independent, despite their conditional independence and the independence of variables $(Z_i)_{i\in[n]}$. Thus the situation is different from the case of the hidden Markov chain models for example, where the observations are not independent because the sequence $(Z_i)_{i\in[n]}$ is a Markov chain, see right-hand side of Figure 4.

Some subvectors of $X$ of a random graph with latent space are nevertheless independent: if $\{i_1, j_1\}, \ldots, \{i_r, j_r\}$ are disjoint pairs, then variables $(X_{i_1 j_1}, \ldots, X_{i_r j_r})$ are mutually independent. Thus, the distribution of each edge $X_{ij}$ conditionally on $Z$ depends only on $Z_i$ and $Z_j$, that is, it involves only the characteristics of the individuals $i$ and $j$ and not anyone alse. Thus the dependency structure is similar to that of Markov models (see paragraph 6): edges are dependent if and only if they share one common node.


 ISSN: 2102-6238

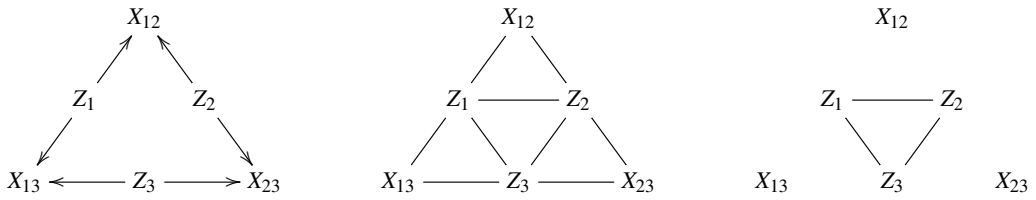


FIGURE 5. *Dependence graph of random graph models with latent variables for $n = 3$ (left), moral graph (middle), dependence graph of $Z$ conditionally on $X$ (right).*

*Dependences of $Z \mid X$: why parametric inference is not possible with the EM algorithm*

There is another consequence of the fact that each observation $X_{ij}$ involves both individuals $i$ and $j$. As Figure 5 shows, the dependence graph of the latent variables $Z$ conditionally on $X$ is a clique, which means that the dependence structure of the conditional distribution cannot be more complex: every latent variable depends on all others. In the parametric case, this complexity prevents from estimating the model with the classical EM algorithm (see more details about this issue in Paragraph 7.5).

### *7.4. First example: the Stochastic Blockmodel (SBM)*

The Stochastic Blockmodel (SBM) corresponds to the simplest case of random graph with latent space, that is, when the set $\mathscr{S}$ is finite; let us define $\mathscr{S} = \{1, \dots, Q\}$, denoted by $[Q]$. Under this model, each node is thus member of a random class/block among $Q$ possible ones. For all $q \in [Q]$, the probability of any node being in class $q$ is denoted by $\alpha_q$. $\alpha$ denotes the probability vector $(\alpha_1, \dots, \alpha_Q)$. Finally, conditionally on $Z$ the probability of two nodes $i$ and $j$ being connected depends only on their classes $Z_i$ and $Z_j$. For all $q, r \in [Q]$, $\pi_{qr}$ denotes the probability of any node from class $q$ and any node from class $r$ being connected, and $\pi = (\pi_{qr})_{q,r\in[Q]}$, called connectivity matrix.

**Latent variables** $(Z_i)_{i\in[n]}$ i.i.d. $\sim \nu = \mathscr{M}(1, \alpha)$

**Observed graph** $(X_{ij})_{i,j\in[n]}$ independent conditionally on $Z$ and for all $i \neq j \in [n]$,
$P(X_{ij} = 1 \mid Z) = \kappa(Z_i, Z_j) = \pi_{Z_i,Z_j}$.

This model is very useful in applied sciences, for instance sociology and biology. In sociology, the SBM is the mathematical incarnation of the principle of social equivalence (Lorrain and White, 1971): any two individuals from the same class share the same probabilities of connection to the other classes and hence play the same social role in the population. For example the popular example of the Zachary Karate Club (Zachary, 1977) illustrates this idea. It is composed of four classes (Léger et al., 2013) highlighted with distinct colors on Figure 6: one class composed of one green node (the president of the club), another one composed of one red node (the karate professor), and one for each of the communities related to these two people: blue for that of the president, mauve for that of the professor.

The parameters are easily interpretable: $\alpha$ gives the proportion of each social group in the whole population, and entries of $\pi$ how socially close are two groups (Holland et al., 1983). In


 ISSN: 2102-6238

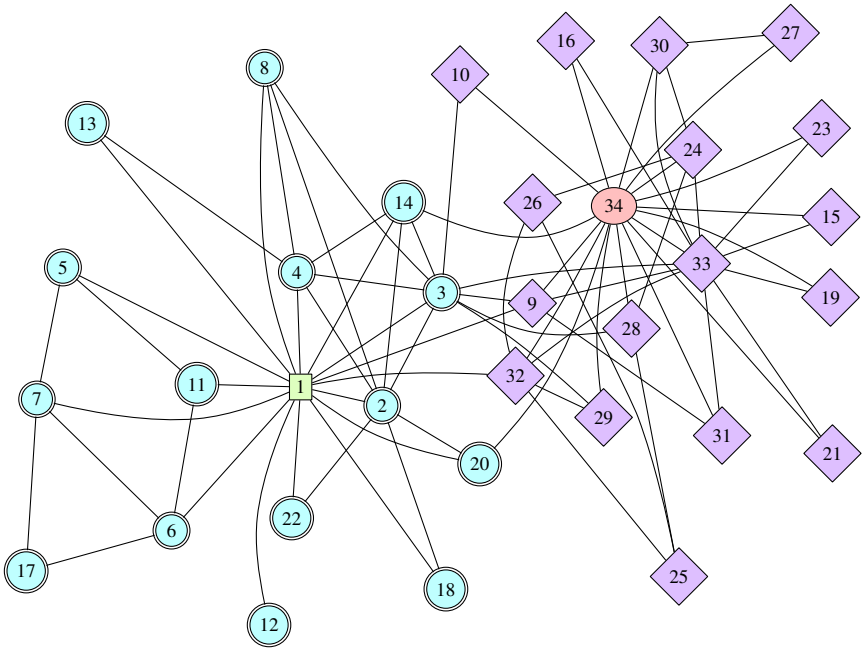


FIGURE 6. *Social network of the Zachary Karate Club*

biology, the SBM is used in particular to model protein-protein interaction networks. Each node symbolizes a protein and the presence of an edge, the existence of a biochemical reaction between the proteins.

One major interest of this model is the concept of *meta-graph*, which is the weighted graph defined by the connectivity matrix $\pi$. Estimating the parameter $\pi$ thus allows to visualize the network on the class-scale instead of the node-scale, which gives a synoptic overview of the network structure. It especially helps to understand deeply the social organization of a population in sociology, or the auto-regulation system of the proteins in biology (see an example of application in Picard et al., 2009).

### *7.5. Inference of the SBM*

Under the Stochastic Blockmodel and some other parametric random graph models with latent variables, the inference is not standard: direct maximization of the observed log-likelihood fails because of its computational complexity, which is largely shared with most models with missing variables, not necessarily in the context of graphs. In the following paragraphs, we show the angles of the inference under the SBM: from why in the context of graphs EM algorithm does not help unlike many other types of data with latent variables, to the approximate inference with the variational formulation of the EM algorithm.


 ISSN: 2102-6238

*Failure of the direct maximum-likelihood approach*

$\theta$ being the parameter of the distribution, two notions of log-likelihoods can be defined in the context of latent variables: the observed (or incomplete) log-likelihood, that is, the log-likelihood of the observed variables $X$, defined as $\mathscr{L}(x,\theta) = \log(P_\theta(X = x))$, and the complete likelihood, that is, the log-likelihood of all variables, whatever their status is, observed or latent: $\mathscr{L}(x,z,\theta) = \log(P_\theta(X = x, Z = z))$.

The observed log-likelihood only can be calculated from the observed variables:$\mathscr{L}(x,\theta) = \log \sum_{z\in[Q]^n} P_\theta(X = x, Z = z)$. The direct maximization is generally not feasible from the point of view of combinatorics, as it requires to explore the set of the possible configurations of the labels. In the Stochastic Blockmodel for instance, the size of this set is $Q^n$: it turns out to be impossible to proceed beyond a few tens of nodes in practice. The complete log-likelihood is easier to calculate but it depends on the latent variables and cannot be hence used directly.

*Difficulties of the EM algorithm*

In the context of models with latent variables other than random graph models, the Expectation-Minimization algorithm (EM algorithm), could appear as a crumb of hope: it has been largely used since late seventies and it is still often used, as the algorithm is simple to run, and the method provides explicit estimators in models of the exponential family in particular.

Define the expectation of the complete log-likelihood at $\theta$, conditionally on the observation and the parameter $\theta'$: $Q_{\theta',X}(\theta) = \mathbb{E}_{\theta'}(\mathscr{L}(X,Z,\theta) \mid X)$, where $\mathbb{E}_{\theta'}$ is the expectation under the distribution of $Z$ associated with the parameter $\theta'$. Starting from an initialization value of the parameter $\theta$, EM algorithm consists in running iteratively the two following steps, until some stopping condition is reached. At iteration $k$:

**E-step** Computing $Q_{\theta^{(k)},X}(\theta)$ for all $\theta$.

**M-step** Maximizing $Q_{\theta^{(k)},X}(\theta)$ with respect to $\theta$, and updating current parameter

$$\theta^{(k+1)} = \arg\max_\theta \left\{ Q_{\theta^{(k)},X}(\theta) \right\}.$$

This strategy, developed in the founding article Dempster et al. (1977), makes indirectly grow the observed log-likelihood at each iteration.

In practice, EM algorithm requires the distribution of $Z$ conditionally on $X$ to be tractable. Mixture models for individual data provide the perfect configuration, since latent variables $Z$ are independent conditionally on observed ones $X$: their distribution is hence factorizable which allows the computation of the expectation at the E-step. On the contrary, since the dependence structure of the distribution of $Z \mid X$ is the worst possible (see Paragraph 7.3), no factorization is possible and E-step is therefore as hard as the direct maximization of the log-likelihood. As a conclusion, the classical EM algorithm does not help at all in the context of random graph models. However, as for the inference under the ERGM (see Paragraph 6.3), dependences can be ignored to approximate nevertheless the conditional distribution of $Z \mid X$ by a product. This idea has been successfully applied through the variational formulation of the EM algorithm.


 ISSN: 2102-6238

*Approximate inference via the variational approach*

The E-step of the EM algorithm amounts to solving a variational problem, that is, an optimization problem over some distribution class. The observed log-likelihood can be first decomposed as follows, for all possible distribution $R$:

$$\mathscr{L}(x,\theta) = \mathscr{F}_{x,\theta}(R) + D_{KL}(R||P_\theta(Z|X=x))$$

where $\mathscr{F}_{x,\theta}(R) = \mathbb{E}_R\left(\log\left(\frac{P_\theta(X=x,Z)}{R}\right)\right)$ and $D_{KL}$ is the Kullback-Leibler divergence. $\mathscr{F}_{x,\theta}$ is the criterion to maximize, indeed:

$$\mathscr{F}_{x,\theta}(R) \leq \mathscr{L}(x,\theta)$$

with equality if and only if distribution $R$ is the real distribution of $Z$ conditionally on $\{X=x\}$. It is therefore the unique maximizer of this criterion over the set of all possible distributions. The key of the variational approach consists in choosing a smaller and nicer class of distributions, over which the maximization problem can be analytically and/or computationally solved.

Like for the classical EM algorithm, two steps are iteratively processed. At iteration $k$:

**VE-step** Find the maximizer $R^{(k)}$ of $\mathscr{F}_{X,\theta^{(k)}}$ among the chosen class of distributions (solve the variational problem).

**VM-step** Find the maximizer $\theta^{(k+1)}$ of $\mathscr{F}_{X,\theta}(R^{(k)})$ with respect to $\theta$.

This algorithm has been successfully adapted to the SBM in Daudin et al. (2008), and has been implemented in the Wmixnet package, described in Léger (2014).

*Dependence structure of $Z \mid X$ and approximate inference*

Generally, and in the present article, the variational problem is solved over the class of product-distributions. Whenever the real conditional distribution is factorizable, the solution is exact and the variational formulation is completely equivalent to that of classical EM. Otherwise, this solution is an approximation of the real conditional distribution and the algorithm provides approximate estimators at the VM-step. Note that in physics, this approximation is also called mean-field approximation (Opper and Saad, 2001).

The factorizability actually depends on the dependency structure of the real distribution of $Z$ conditionally on $X$. For example in finite mixture models for individual data, variables $(Z_i)_{i\in[n]}$ are independent conditionally on $X$ and their distribution is thus factorizable. More generally, whenever the dependence graph is a tree[9], the distribution is always factorizable and a sum-product algorithm achieves the analytical computation of the exact solution of the variational problem (the real conditional distribution). No approximation is actually made in these cases. On the contrary, the structure dependence is a clique in the setting of random graphs (see Paragraph 5). The conditional distribution of $Z$ conditionally on $X$ cannot be factorized at all. Solving the variational problem over the set of all possible distributions would be as hard as the direct maximization of the observed log-likelihood or as the classical EM. In this very situation it is

[9] For example, outside the field of networks, in models with a hidden Markov chain (HMM), the conditional distribution is factorizable step by step, owing to the unidimensional Markovian dependence structure: the dependence graph of $Z$ conditionally on $X$ is a wire (Cappé et al., 2006).


 ISSN: 2102-6238

ideal that the mean-field approximation comes into play. For instance in the SBM, solving the variational problem over the class of factorized distributions is feasible: it is equivalent to solving a fixed point equation (Daudin et al., 2008). Then the VM-step provides approximate estimators of $(\alpha, \pi)$.

Properties of mean-field approximation are not fully understood yet. The method is already known to be not consistent in general (Wang and Titterington, 2004). However, it turns out to be consistent under the SBM, which is proved in Célisse et al. (2012). Despite the complexity of its dependence structure, the conditional distribution of $Z$ converges to a Dirac located at the real classification (Mariadassou et al., 2015), which is factorizable. More generally, the method seems to be consistent in models such that the limiting conditional distribution is factorizable, that is, whenever the mean-field approximation is asymptotically exact. It is precisely satisfied by the SBM.

The variational approach has many fruitful developments, including other types of models with latent variables, for instance models with hidden Markov fields (Zhang, 1992), coupled HMM (Murphy, 2002) or in a Bayesian framework (Beal, 2003). See Wainwright and Jordan (2008); Jaakkola (2000) for a general introduction to the variational method.

*Conclusion and alternative methods*

There are two types of inference strategies in models with latent variables. One of them iteratively processes both classification and estimation jointly, the one improving the knowledge of the other at each step. Variational EM algorithm is like this for example. In this vein, note that Snijders and Nowicki (1997) proposed a Bayesian method with a Gibbs sampling in the case of $Q = 2$ classes, generalized in Nowicki and Snijders (2001).

On the contrary, the second strategy consists in first classifying and then estimating the parameters. Indeed, the classification is the most difficult task and once the classification done, it is quite easy to derive estimators of the parameters. These methods are mostly *ad hoc* procedures and depart from the classical maximum-likelihood approach. For example Rohe et al. (2010) proposes a consistent classification algorithm adapted from the spectral clustering (Von Luxburg, 2007; Bhattacharyya and Bickel, 2015). Channarond et al. (2012) also proposes a consistent classification algorithm called LG. Parameter estimators and a model selection criterion are derived from it, as well as a test of heterogeneity, added as an appendix in Channarond (2013).

The variational EM procedure is the most commonly used, as it provides good results, even for small graphs. This method can actually be used only for small graphs up to few thousands of nodes, because it is rather computationally complex. Note that spectral clustering is used as initialization classification in the Wmixnet package mentioned in Paragraph 7.5. Finally LG algorithm is able to process very large graphs, up to millions of nodes. It gives good results in simulation for such graphs.

All inference methods in the context of random graph models share one common feature: they mostly converge very fast, because the number of observations $n(n-1)/2$ increases very fast with respect to the number $n$ of individuals. Each new individual connects (or does not) to the individuals already present, which brings $n-1$ new pieces of information, that is one on each individual. Snijders and Nowicki (1997) and Nowicki and Snijders (2001) notice that as a result, the prior influence vanishes very quickly in their Bayesian approach. Moreover, marginal


 ISSN: 2102-6238

distributions, as the degree distribution, concentrate very fast as well. Under some separability assumption, Channarond et al. (2012) proves that it is possible to carry out the whole inference of the SBM, and thus to answer all questions asked in Paragraph 7.1, only from the degrees. As a conclusion, even though the dependence structure of $Z \mid X$ is complex, the inference of the SBM is theoretically much easier than expected, as a result of the concentration.

### 7.6. *Second example: random graphs with latent positions*

In this subfamily of models, the latent space $(\mathscr{S}, \nu)$ is such that $\mathscr{S}$ is a metric space, equipped with a distance $\delta$. The latent variable of a node can be thought of as a geographical position in the space, or as a set of social characteristics, and hence as a social position; moreover the latent space is sometimes called social space in the literature, see Hoff et al. (2002); Handcock et al. (2007). The distance $\delta$ measures how close or similar are any two nodes in the latent space. Following the idea, emerged in the seventies, that social networks are ruled by social distances (McFarland and Brown, 1973), it is generally assumed that the conditional probability of connection of any two nodes depends only on their distance. Furthermore the closer or the more similar any two nodes are, the larger is their probability of connection, according to the social paradigm " birds of a feather flock together " . In the general framework described previously, the model is defined as follows:

**Latent variables** $(Z_i)_{i\in[n]}$ i.i.d. $\sim \nu$.

**Observed graph** $(X_{ij})_{i,j\in[n]}$ independent conditionally on $Z$ and for all $i \neq j \in [n]$,
$$\kappa(Z_i, Z_j) = P(X_{ij} = 1 \mid Z) = k\left(\delta(Z_i, Z_j)\right).$$

where $k : \mathbb{R}^+ \to [0,1]$ is therefore a non-increasing function.

Random geometric graphs are a very particular example, we define $r > 0$ and $\kappa(z,z') = \mathbb{1}_{\{\delta(z,z')\leq r\}}$ for all $z, z' \in \mathscr{S}$. The graph is therefore deterministic conditionally on $Z$: any two nodes connect if and only if their distance is less than $r$. See Paragraph 4.2 for more details.

In most of the cases, $\mathscr{S}$ is $\mathbb{R}^d$ for some $d \in \mathbb{N}^*$, $\delta$ is the euclidean distance and $\nu$ has some density $f$ with respect to the Lebesgue measure on $\mathbb{R}^d$. In the Latent Cluster Position Model (Handcock et al., 2007), $f$ is a Gaussian mixture and the connection function $\kappa$ is a non-increasing logistic function of the distance. The inference strategy consists in estimating the positions of the nodes up to isometries, and then plugging them into a classical EM for gaussian mixtures. The positions are estimated by first estimating the distances by maximum-likelihood estimation (Hoff et al., 2002), and then using MultiDimensional Scaling (MDS). The authors of Handcock et al. (2007) admit that this method processes the estimation of the positions and of the mixture parameters separately, although they are strongly related. Thus they also proposed a fully integrated Bayesian procedure to overcome this default. The different steps of these inference methods are standard, but their piling makes the theoretical analysis of the global procedure hard. Moreover, they use iterative algorithms which are computationally costly, and are therefore not able to process very large graphs.

### 7.7. *Discussion about clustering and random graph models with latent variables*

Let us stress once again that clustering is a vague notion. The latent space $(\mathscr{S}, \nu)$ shapes the type of clusters, which will be obtained by the inference methods. Their choice is thus critical and


 ISSN: 2102-6238

should be adapted to the objectives of the statistical analysis and driven by the application.

The SBM enables to obtain varied types of clusters going from well-connected communities to groups of individuals sharing the same hierarchical level for instance (see examples in Daudin et al., 2008). However, this model can be considered as rigid since the spectrum of social roles is finite, and social roles cannot be combined. Some variants propose to make the structural equivalence more flexible. The Overlapping SBM allows any node to be member of more than one class (Latouche et al., 2011) or possibly none; such nodes are declared as outliers. The Mixed Membership SBM (Airoldi et al., 2008) models the fact that an individual can play several distinct social roles, according to the individual it interacts with. In this case the latent space is the $Q$-dimensional simplex, which is continuous. It is not an exhaustive list of the variants of the SBM, which is quite burgeoning at this time.

Using a continuous latent space allows more flexibility, but clustering is harder to define. The latent variables of the SBM are directly classes indicating to which cluster nodes belong to, whereas rules defining clusters are needed in a continuous setting. The LPCM proposes to assume the distribution of the latent variables to be a finite gaussian mixture, so that clusters are defined by extra latent variables in a finite space, actually equivalent to the latent classes of the SBM. However this implies some constraints on the geometric shape of the clusters, because the distribution of the latent variables is parametric. Thus it seems desirable to make as few assumptions as possible about the density $f$ of the positions $Z$, in particular with the ultimate aim of constructing a robust test of heterogeneity (first challenge mentioned in Paragraph 7.1). Channarond (2013) proposes a non-parametric setting, and defines the clusters geometrically with the level sets of the density $f$ (Hartigan, 1975).

### *7.8. Third example: graphon and justification of the random graphs with latent space*

**Definition 7.3.** *A random graph $X$ is said to be a graphon if it is a random graph with latent space $(\mathscr{S}, \nu)$ and connection function $\kappa$, such that $\mathscr{S} = [0,1]$, $\nu$ is the uniform distribution over $[0,1]$ and $\kappa : [0,1]^2 \to [0,1]$ is a symmetric map.*

The distribution of such a random graph depends only on the number of nodes $n$ and the symmetric map $\kappa$, and is denoted by $\mathscr{H}_{n,\kappa}$. For all $n \in \mathbb{N}^*$ and all symmetric measurable map $\kappa : [0,1]^2 \to [0,1]$, the definition essentially describes the following random graph:

**Latent variables** $(Z_i)_{i\in[n]}$ i.i.d. $\sim \mathscr{U}([0,1])$

**Observed graph** $(X_{ij})_{i,j\in[n]}$ independent conditionally on $Z$ and for all $i \neq j \in [n]$,
$$P(X_{ij} = 1 \mid Z) = \kappa(Z_i, Z_j).$$

The reason of our interest in graphons, which is just a subfamily of random graphs with latent space, is the representation theorem from Lovász and Szegedy (2006) (Theorem 2.7 in this reference): a large class of random graphs, which can be sequentially constructed, are actually graphons.

**Theorem 7.1.** *(Lovász and Szegedy, 2006) Let $(X^n)_{1\leq n\leq N}$ be a random graph sequence such that for all $n \in [N]$, $X^n$ has $n$ nodes. There exists a symmetric measurable map $\kappa : [0,1]^2 \to [0,1]$ such that for all $n \in [N]$ the distribution of $X^n$ is $\mathscr{H}_{n,\kappa}$ (hence $X^n$ is a graphon) if and only if*

1. *the distribution of $X^n$ is invariant up to node relabeling,*


 ISSN: 2102-6238

2. *for all* $n \in \{2,\dots,N\}$*, the distribution of the subgraph of* $X^n$ *induced by the node set* $[n-1]$ *is the same as the distribution of* $X^{n-1}$*,*
3. *for all* $1 < n < N-1$*, subgraphs of* $X^N$ *induced by the node sets* $\{1,\dots,n\}$ *and* $\{n+1,\dots,N\}$ *are independent.*

Every random graph with latent space is hence a graphon. For example, every random graph from the SBM with $n$ nodes whose parameters are $(\alpha,\pi)$ is the graphon $\mathscr{H}_{n,\kappa}$ with $\kappa$ a piecewise function defined as follows. $[0,1]^2$ is split into $Q^2$ disjoint rectangles; each one corresponding to one couple of classes. For the rectangle associated with classes $(q,r) \in [Q]^2$, the lengths of the sides are $\alpha_q$ and $\alpha_r$ and the value of the function on this rectangle is $\pi_{qr}$.

However the theorem is only theoretical in general, and does not construct an explicit map $\kappa : [0,1]^2 \to [0,1]$, which may be possibly very complex and with no regularity. Even in the model with latent positions in $\mathbb{R}^d$ for instance, there exists such a function; nevertheless it is much more complicated to provide it. Moreover, note that the asymptotic framework where the graph latent space depends on $n$ does not satisfy the assumptions of the representation theorem.

Note that the graphon model is not identifiable: any isometry of the unit square changes the graphon function $\kappa$, but does not change the distribution of the random graph. Up to some equivalence between random graph distributions, $\kappa$ can be nevertheless estimated, in particular by approximating it by a piecewise function (Airoldi et al., 2013; Olhede and Wolfe, 2014), which actually amounts to approximate the graphon model by a Stochastic Blockmodel, as explained previously. See also Chatterjee et al. (2014), which provides an estimator more generally designed for structured matrices.

## References


Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014.

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

Albert, R. and Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Aldous, D. (1997). Brownian excursions, critical random graphs and the multiplicative coalescent. *The Annals of Probability*, pages 812–854.

Arias-Castro, E. and Verzelen, N. (2013). Community detection in random networks. *arXiv preprint arXiv:1302.7099*.

Arratia, R., Goldstein, L., and Gordon, L. (1989). Two moments suffice for Poisson approximations: the Chen-Stein method. *The Annals of Probability*, 17(1):9–25.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Barbillon, P., Thomas, M., Goldringer, I., Hospital, F., and Robin, S. (2015). Network impact on persistence in a finite population dynamic diffusion model: application to an emergent seed exchange network. *Journal of theoretical biology*, 365:365–376.

Barbour, A. and Mollison, D. (1990). Epidemics and random graphs. *Stochastic processes in epidemic theory*, 86:86–89.

Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson approximation*. Clarendon press Oxford.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London.

Bender, E. A. and Canfield, E. R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307.

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

Besag, J. (1975). Statistical analysis of non-lattice data. *The statistician*, pages 179–195.

Besag, J. (2001). Markov chain Monte Carlo for statistical inference. *Center for Statistics and the Social Sciences*.



 ISSN: 2102-6238

Bhattacharyya, S. and Bickel, P. J. (2015). Spectral Clustering and Block Models: A Review And A New Algorithm. *arXiv preprint arXiv:1508.01819*.

Birmele, E. (2012). Detecting local network motifs. *Electronic Journal of Statistics*, 6:908–933.

Blondel, V., Krings, G., and Thomas, I. (2010). Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies*, 42(4).

Bollobás, B. (2001). *Random graphs*. Cambridge Univ Pr.

Bollobás, B., Janson, S., and Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122.

Bollobás, B. and Riordan, O. (2003). Mathematical results on scale-free random graphs. *Handbook of graphs and networks*, 1:34.

Bollobás, B. and Riordan, O. (2004a). Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*, 1(1):1–35.

Bollobás, B. and Riordan, O. (2004b). The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34.

Bollobás, B., Riordan, O., Spencer, J., Tusnády, G., et al. (2001). The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290.

Buckley, P. G. and Osthus, D. (2004). Popularity based random graph models leading to a scale-free degree sequence. *Discrete Mathematics*, 282(1):53–68.

Callaway, D. S., Newman, M. E., Strogatz, S. H., and Watts, D. J. (2000). Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25):5468.

Cappé, O., Moulines, E., and Rydén, T. (2006). *Inference in hidden Markov models*. Springer Science & Business Media.

Célisse, A., Daudin, J.-J., Pierre, L., et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

Channarond, A. (2013). *Recherche de structure dans un graphe aléatoire: modèles à espace latent*. PhD thesis, Université Paris Sud-Paris XI.

Channarond, A., Daudin, J.-J., and Robin, S. (2012). Classification and estimation in the Stochastic Blockmodel based on the empirical degrees. *Electronic Journal of Statistics*, 6:2574–2601.

Charbonnier, C., Chiquet, J., and Ambroise, C. (2010). Weighted-LASSO for structured network inference from time course data. *Statistical applications in genetics and molecular biology*, 9(1).

Chatterjee, S. et al. (2014). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.

Chen, L. H. (1975). Poisson approximation for dependent trials. *The Annals of Probability*, pages 534–545.

Chung, F., Lu, L., Dewey, T. G., and Galas, D. J. (2003). Duplication models for biological networks. *Journal of Computational Biology*, 10(5):677–687.

Daudin, J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and computing*, 18(2):173–183.

Dempster, A., Laird, N., Rubin, D., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dommers, S., van der Hofstad, R., and Hooghiemstra, G. (2010). Diameters in preferential attachment models. *Journal of Statistical Physics*, 139(1):72–107.

Draief, M. and Massoulié, L. (2010). Epidemics and rumours in complex networks, volume 369 of London Mathematical Society Lecture Notes.

Durkheim, E. (1893). *De la division du travail social: étude sur l'organisation des sociétés supérieures*. F. Alcan.

Durrett, R. (2007). *Random graph dynamics*, volume 20. Cambridge university press.

Erdős, P. and Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297.

Erdős, P. and Renyi, A. (1961). On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1):261–267.

Erdos, P. and Rényi, A. (1961). On the evolution of random graphs. *Bull. Inst. Internat. Statist*, 38(4):343–347.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

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

Gerbaud, A. (2010). *Modélisation de réseaux d'interactions par des graphes aléatoires*. PhD thesis, Université de Grenoble.


 ISSN: 2102-6238

Gilbert, E. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.
Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233.
Gupta, P. and Kumar, P. R. (1998). Critical power for asymptotic connectivity in wireless networks. In *Stochastic analysis, control, optimization and applications*, pages 547–566. Springer.
Haenggi, M., Andrews, J. G., Baccelli, F., Dousse, O., and Franceschetti, M. (2009). Stochastic geometry and random graphs for the analysis and design of wireless networks. *Selected Areas in Communications, IEEE Journal on*, 27(7):1029–1046.
Handcock, M., Raftery, A., and Tantrum, J. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.
Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.
Hoff, P., Raftery, A., and Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.
Holland, P., Laskey, K., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.
Holland, P. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50.
Hunter, D. R. and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583.
Jaakkola, T. (2000). Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, pages 129–159.
Janson, S., Knuth, D. E., Łuczak, T., and Pittel, B. (1993). The birth of the giant component. *Random Structures & Algorithms*, 4(3):233–358.
Janson, S., Luczak, T., and Kolchin, V. (2000). *Random graphs*. Cambridge Univ Press.
Kleinberg, J. (2000). The small-world phenomenon: an algorithm perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM.
Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The web as a graph: Measurements, models, and methods. In *Computing and combinatorics*, pages 1–17. Springer.
Kolaczyk, E. D. (2009). *Statistical analysis of network data*. Springer.
Latouche, P., Birmelé, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336.
Léger, J.-B. (2014). Wmixnet: Software for clustering the nodes of binary and valued graphs using the stochastic block model. *arXiv preprint arXiv:1402.3410*.
Léger, J.-B., Vacher, C., and Daudin, J.-J. (2013). Detection of structurally homogeneous subsets in graphs. *Statistics and Computing*, pages 1–18.
Lorrain, F. and White, H. (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1):49–80.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*.
Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957.
Luce, R. D. (1952). Two decomposition theorems for a class of finite oriented graphs. *American Journal of Mathematics*, pages 701–722.
Mariadassou, M., Matias, C., et al. (2015). Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573.
Matias, C. and Robin, S. (2014). Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proceedings and Surveys*, 47:55–74.
McFarland, D. D. and Brown, D. J. (1973). Social Distance as Metric: A Systematic Introduction to Smallest Space Analysis. *EO Laumann. Bonds of Pluralism: The Form and Substance of Urban Social Networks. New York: John Wiley*, pages 213–252.
Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.
Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.
Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180.
Moreno, J. L. (1937). Sociometry in relation to other social sciences. *Sociometry*, 1(1/2):206–219.


 ISSN: 2102-6238

Moreno, J. L. and Jennings, H. H. (1938). Statistics of social configurations. *Sociometry*, pages 342–374.

Murphy, K. P. (2002). *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley.

Nachmias, A. and Peres, Y. (2010). The critical random graph, with martingales. *Israel Journal of Mathematics*, 176(1):29–41.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E. and Watts, D. J. (1999a). Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4):341–346.

Newman, M. E. and Watts, D. J. (1999b). Scaling and percolation in the small-world network model. *Physical Review E*, 60(6):7332.

Nielsen, T. D. and Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.

Nowicki, K. and Snijders, T. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Olhede, S. C. and Wolfe, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727.

Opper, M. and Saad, D. (2001). *Advanced mean field methods: Theory and practice*. MIT press.

Penrose, M. (2003). *Random geometric graphs*, volume 5. Oxford University Press Oxford.

Picard, F., Miele, V., Daudin, J., Cottret, L., and Robin, S. (2009). Deciphering the connectivity structure of biological networks using MixNet. *BMC bioinformatics*, 10(Suppl 6):S17.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (< i> p< sup>*</sup>) models for social networks. *Social networks*, 29(2):173–191.

Rohe, K., Chatterjee, S., and Yu, B. (2010). Spectral clustering and the high-dimensional Stochastic Block Model. *Arxiv preprint arXiv:1007.1684*.

Snijders, T. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100.

Snijders, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Snijders, T. A. (2011). Statistical models for social networks. *Annual Review of Sociology*, 37:131–153.

Snijders, T. A., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153.

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

Van Der Hofstad, R. (2009). Random graphs and complex networks. *Available on http://www. win. tue. nl/rhofstad/NotesRGCN. pdf*.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

Wang, B. and Titterington, D. (2004). Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, 20(3):151–170.

Wasserman, S. and Robins, G. L. (2005). An introduction to random graphs, dependence graphs, and p*. *Models and methods in social network analysis*, 27:148–161.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442.

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473.

Zhang, J. (1992). The mean field theory in EM procedures for Markov random fields. *Signal Processing, IEEE Transactions on*, 40(10):2570–2583.


 ISSN: 2102-6238