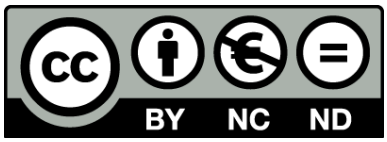
UNIVERSITAT DE BARCELONA

The role of clustering in the stucture and function of complex networks

Pol Colomer de Simón



UNIVERSITAT DE BARCELONA

PHD THESIS

PROGRAMA DE DOCTORAT EN FÍSICA

The role of clustering in the stucture and function of complex networks

Author:
Pol COLOMER DE SIMÓN

Supervisor:
Dr. Marián BOGUÑÁ

Tutor:
Dr. Albert DÍAZ GUILERA

UNIVERSITAT DE BARCELONA

Acknowledgements

A PhD is commonly conceived as a hard and solitary period, and in some aspects it is true. However, I have been very lucky to have many people around me that gave me scientific and personal support to overcome the multitude of problems that I have faced during this long process. In this section, I would like to thank all those people that, in one way or another, contributed to the work I am presenting here.

In the first place, I would like to thank my PhD supervisor and principal responsible that I have been able to come this far: Marián Boguñá. Per començar, gràcies per des d'un principi fer-me confiança oferint-me una beca sense la qual segurament no hagués acabat fent estudis de doctorat. Espero no haver-te defraudat. Però sobretot, gràcies per ensenyar-me un munt de coses i per animar-me enlloc de pressionar-me quan les coses no sortien com esperàvem. Sempre m'he sentit molt afortunat de tenir un director de tesi amb el teu grau de implicació, vocació didàctica i nivell científic sempre a la meva disposicó. Professionalment t'ho dec tot.

I also want to thank Dmitri Krioukov who made my stay at the UCSD comfortable, pleasant and enricher. Thanks Dima, for your hospitality and always interesting conversations. I consider you a friend and I hope to see you more times in the future.

Thanks also to all members of the Complex Lab Barcelona research group at the university of Barcelona: Albert, María Ángeles, Conrad, Josep, Oleguer, Nikos, Antoine, Kolja, Oriol, Roberta, Guille, Mario, Rubén, Ignacio, Luche, Jan, Francesco and Michele. Working in such a healthy human atmosphere made it very easy to keep the motivation to go to work everyday. Among the members of the group, I would like to make an special mention to whom I consider my PhD soulmate: Oleguer Sagarra. Sort de les incomptables riures que ens hem pegat aquests quatre anys al despatx, als dinars, als congressos... Realment sense tu això hagués sigut infumable. Encara que siguis un petat espero poder compartir més projectes i experiències il·lusionants amb tu. Gràcies especials també a l'Albert Díaz per generar bones dinàmiques de treball dintre el grup i tractar-me i cuidar-me com si fos el seu estudiant.

Querría agradecer también a Mariano Beiró i Ignacio Álvarez Hamelin por la enriquecedora colaboración en el desarrollo de la herramienta de visualización LaNet-Vi 3.0.

Also thanks to Ali Faqeeh, Sergey Melnik and James P. Gleeson for their hospitality and interesting discussions during my short stay in Limerick .

També és important donar les gràcies a tots amb els qui he compartit despatx durant aquests 4 anys: Adriana, Blai, Carlos, Genís, Eloi, Oleguer, Kolja, Mario i

Roberta. Una especial menció es mereix L'Oriol Vilanova. Mai havia conegut ningú que li importés tant poc deixar de fer el que un està fent per ajudar al company. Em sap greu no haver-te pogut ajudar tant com tu m'has ajudat a mi.

Gràcies també a les secretàries del departament, l'Olga, la Bea, La Cristina i l'Elena. No sóc l'únic del departament que dóna les gràcies cada dia per tenir un equip de secretaria tant eficient. Demano perdó si algun cop us he donat feina extra per perdre tiquets, cartes d'embarcament o he oblidat posar el NIF de la UB en alguna factura.

I also want to thanks Chiara Orsini and Massimo Ostilli who, besides the interesting scientific interactions, made my stay in the UCSD very easy and amusing.

Thanks also to Chiara Orsini, Sergio Oller, Oriol Vilanova for their contribution to the development of the random network generator code RandNetGen.

També una menció molt especial al grup de persones amb qui he compartit la major part dels dinars, molt importants cara a recarregar piles per afrontar les tardes. Una menció especial per als assistents més recurrents: Núria, Dani, Enric, Sergio, Eloi, Ori. Sens dubte el dinar ha sigut el millor moment del dia d'aquests quatre anys. Motiu suficient i de sobres per venir a treballar a la universitat cada dia. Sort he tingut de vosaltres. Gràcies en especial a la Núria per impregnar els dinars, les pauses i el pis amb la teva alegria.

També voldria agrair a tots aquells que hem coincidit en congressos, workshops i de més. Persones com la Toñi, el Toni, el Dani, el Fran i altres que han fet que tots aquests esdeveniments, a més d'interessants, fossin també divertits.

Gràcies també a tots els components de Físics pel món. Ja són 10 anys a la facultat i com sempre el més important que m'emporto són les persones que he conegut. Que bé m'ho vaig passar a la carrera sobretot gràcies al Gual i a l'Héctor. Gràcies també Clotet per el teu mail sobre tots els requeriments burocràtics per el dipòsit de la tesi.

També gràcies a la Mar Güell amb qui he rigut molt i m'ha fet millor persona.

Un cop has passat una setmana dura de feina no hi ha res millor que tenir per davant un cap de setmana ple de diversió. D'aquesta part bàsicament s'han encarregat els reclutes: Hug, Max, Estruch, Keke, Keis, Bea, Jaume, Eudald, Rai, Andreu i Pol. Gràcies per ser sempre allà. Sabeu de sobres com us estimo i us necessito. Menció especial al keke per ajudar-me en el disseny de la portada.

Gràcies a L'Anna per alimentar-me com Déu mana i a la Tata, la dona més forta que he conegut mai que m'ha cuidat i estimat com si fos el seu fill. Et trobaré molt a faltar.

L'últim agraïment el dedico a tota la meva família. El vostre amor incondicional ha sigut bàsic per poder tirar endavant aquest projecte. En especial, gràcies als meus pares a qui bàsicament els ho dec tot. Gràcies, us estimo.

Contents

CHAPTER 1

Introduction

1.1 Network science

In the sense in which I will use it here, a "network" or "graph" is an abstract mathematical entity consisting of points or nodes connected by edges. The structure of such a network is so simple that it can be used to represent any system with interacting elements. Examples of such systems can be found all around us and can have very different natures. Perhaps the most immediately obvious example are social networks, in which individuals are related by different types of social interactions: acquaintance, sexual contact, Facebook friendship, twitter following [81, 111], etc. There are also technological networks such as the Internet, in which we have routers, devices that share data through physical wires [146]; or power grids, in which we have generating stations and switching substations connected by high-voltage lines [170]. There are also biological networks such as the brain, made of neurons connected by their axon terminals where synapses occur; or food webs, in which the species that comprise an ecosystem are related by who eats whom [63].

With all theses systems, considering their network structure as a whole, instead of studying the interactions among the elements independently, can have a striking effect when it comes to making decisions related to them. For instance, in the 80's, there was a dramatic reduction of the north-west Atlantic stock of cod that resulted in an economic crisis in the Canadian fishing industry. As a response, the Canadian government financed massive seal culls arguing that a reduction in the numbers of the major predator of cod would increase the cod stock. After the slaughter of nearly half a million seals, the cod population continued to decline. Reconstruction of the north-west Atlantic food web showed that among the species preyed on by seals, were many other predators of cod. These indirect relations between cod and seals suggested that the decrease of the seal population could ultimately have an effect on the cod stock that was the opposite of initial expectations [75].

In other situations, the unpredictable effect that such complex connectivity patterns have on network function can be extremely important. For example, on the night of November 4th 2006, the German electricity company E.ON switched off an electricity line across the River Ems to allow a cruise ship to pass along

the river safely. The removal of just that one connection triggered a cascade of failures that left some 15 million households without electricity across Germany, France, Italy, Spain and Portugal [113].

Apart from ecological balance and the collapse of power grids, other examples of phenomena that occur in networks that are still beyond our knowledge include traffic jams in cities, the spread of infectious diseases, so-called "virality" within on-line social networks and epilepsy attacks. Such phenomena are hard to predict not because we do not understand the nature of each element of the system, but due to emergent properties generated by the complex patterns of interaction between them. So there is a need for new theories that reveal the effect that the connectivity of a system has on its behaviour, thereby bridging the gap between network structure and function. With this purpose in mind, network science has provided many tools that bring us closer to an accurate understanding of such phenomena [138].

The foundations of network science were laid in the 18th century. In 1735, Leonhard Euler tried to find a route that crossed each of the seven bridges of the city of Königsberg (currently Kaliningrad, Russia) only once. To solve this problem, Euler created an abstract mathematical structure in which land masses and bridges were substituted by nodes and edges respectively; constructing what is known as a graph or a network (see Fig. 1.1). From this perspective, Euler observed that since one must both enter and leave every node within the route except the first and the last, there can be at most two nodes with an odd number of connections [68]. However, all the landmasses in the network had an odd number of bridges. So, Euler not only proved that the bridge problem had no solution, but stated the first theorem of graph theory, which became a new framework for the study of systems in which connectivity between the elements plays a prominent role in their behaviour.

In the beginning, graph theory only attracted the interest of mathematicians; they studied the topology of simple graphs such as lattices, trees and random graphs [66, 163]. In the mid 20th century, sociologists and anthropologist became interested in the idea of social networks and they applied some of the notions of graph theory to them. For instance, in 1958, Sola Pool and Manfred Kochen were interested in questions such as: What is the distribution of the number of acquaintances that people have? Who are the most influential people in the network? How far apart are any two people chosen at random? What is the exact structure of the network? At that time, access to empirical data on social networks was very limited because it only relied on interviews, which were expensive, restricted and unreliable. Nevertheless, to address these questions, sociologists applied random graph theory and made some conjectures. For instance, they claimed that the world was becoming smaller, so any two individuals could be connected via a smaller number of social contacts [57].

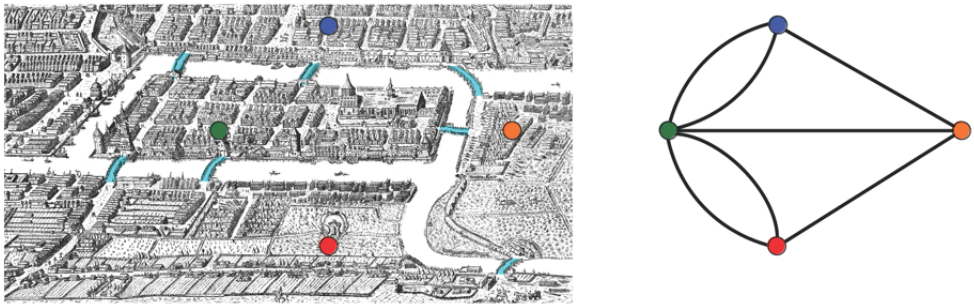


Figure 1.1: The city of Königsberg with its seven bridges. In 1735 Euler used a network representation to prove that there was no route that crossed each of the seven bridges of the city of Königsberg only once.

Despite the methodological difficulties, much systematic recording and analysis of social interactions was performed. In 1967, Stanley Milgram devised a rigorous experiment to track chains of acquaintances and reported the first empirical evidence of the so-called *small world* property of social networks [168]. In his experiment, Milgram selected a set of individuals who had to deliver a letter to a specific person by passing it to one of their acquaintances with a high probability of knowing the target individual. Those acquaintances were asked to do the same. Out of 64 letters that reached their destination, the average number of stages or "hops" in the social network to reach the target person was 6.2, giving rise to the popular concept of "six degrees of separation".

At that time, applications of graph theory were restricted to social networks and still not much was known about their structure. However, the introduction of electronic databases, the increase in the use and power of computers, and the propagation of the Internet dramatically increased access to large network datasets. As a consequence, many different studies analysing the structure of technological [1, 69], social [130] and biological networks [96] were published. Those datasets were far larger than the results of traditional sociometry, so it was much more appropriate to apply to them the statistical tools that physicists and mathematicians were familiar with. Surprisingly, all the work revealed that the small world property is not exclusive of social networks, but a common rule in all interconnected systems. Moreover, other structural features were repeatedly found within networks that had completely different natures [34, 170]. The similarities among these complex architectures, which are neither purely regular nor purely random, suggested the presence of common formation mechanisms and stirred interest in network science among scientists from many different areas.

The most important common feature in many networks is the lack of a typi-

cal scale of the number of connections per node. This *scale free* property implies that a small fraction of the nodes drive the behaviour of the whole system: they are the hubs. This topological property of real networks was first discovered in the network of scientific papers by Derek De Solla Price in 1965 [58] and then in a wide variety of other networks [2, 69, 96]. Subsequent studies found that this heterogeneity of nodes makes networks more resilient and more navigable; but at the same time, more prone to spread infectious agents [24].

Another striking topological feature present in many real networks is the propensity of nodes to cluster together. This tendency of nodes to have many neighbours in common translates into a high concentration of triangles, known as *high clustering*. In social networks, this feature becomes obvious once we observe that of our friends, many are also friends with each other. In this case, it does not seem so surprising to us, as our new friends are usually introduced to us by our existing contacts. In 2002, however, Ron Milo et al. found that the presence of such short loops was also much higher than could be attributed to mere random chance in biological and technological networks [123]. Thus, the surprising constant presence of these regular patterns in all types of networks makes clustering one of the most important factors in the common formation mechanisms of real-world networks.

From that point onwards, network science has undergone many advances and a broad set of tools for analysing and understanding networks have been developed [135]. Nowadays, a large variety of algorithms are available to measure how well connected nodes are or to calculate the shortest path between any pair of nodes; as are a wide range of techniques that group nodes into meaningful communities [72]. There are also network models whose aim is to reproduce network evolution, providing much insight into network formation mechanisms [32]. There are many models that allow us to make predictions regarding how robust power grids are when faced with random failures of their constituent parts or the evolution of a disease within a certain population [13, 165].

The wide applicability of network science has brought together researchers from many different areas including mathematics, physics, biology, epidemiology, computer science and sociology. This fact combined with the huge amounts of data available nowadays makes network science a very fruitful field with a large scientific production and a broad range of real applications. Indeed, ideas from network science have been applied, for instance, to: the analysis of metabolic and genetic regulatory networks [115, 158]; the design of efficient communication protocols, to solve serious scaling limitations that the Internet faces today [28]; the development of vaccination strategies for the control of diseases [145]; and marketing campaigns to increase their success [101].

However, the major contributions to network science have occurred over the last 20 years; so many challenges still lie ahead. Among those challenges, ques-

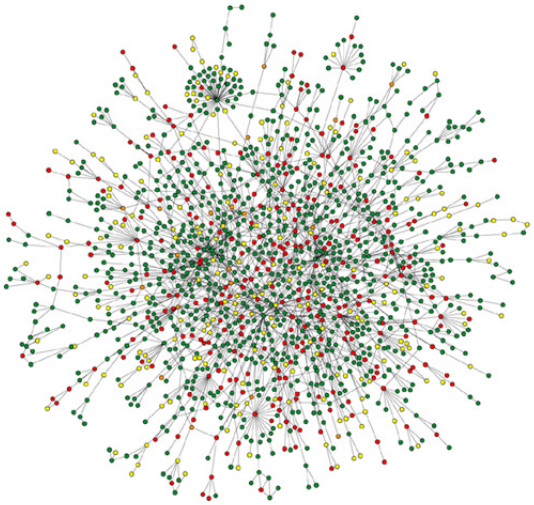


Figure 1.2: A representation of the protein—protein interaction network in *Saccharomyces cerevisiae*, which is based on early yeast two-hybrid screening. The colour of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown) [10]

tions concerning clustering stand out from the rest. In contrast to the small world and scale free properties, the emergence of clustering and its effects on other topological properties and network processes are not fully understood. The reason for this is twofold. First, the mere presence of triangles in networks contradicts assumptions that are commonly used in theoretical approaches of networks. Second, there is a lack of clustered network models that reproduce real network topology to study the effects of clustering on network structure and function. Therefore, clustering is the main factor that hinders the possibility of accurate predictions in real situations; so it constitutes one of the major challenges that network science faces.

This thesis aims to contribute to our understanding of the role played by clustering in the structure and function of complex networks. I show that clustering can have an striking effect on the global structure of networks and that this changes completely how we understand critical network phenomena. Applying this new paradigm to real networks, we are able to grasp the weak points of certain state-of-the-art theories and therefore it contributes to much more accurate understanding of network processes.

1.2 Network topology

From any given dataset that contains a list of the interactions between the elements of a system, we can construct a network. Once we have constructed the network, there are a variety of useful measures that captures the most important features of its topology.

1.2.1 Network representation

Networks can be classified depending on how much information we want to encode in the edges. In a *simple graph* there can only be one connection between any pair of nodes; if there are more, it is a *multigraph*. In some systems, the interaction between elements is not reciprocal, so the edges that represent those interactions have a direction, from one vertex to another. Networks with edges of this kind are *directed networks*. In some situations it is useful to represent the different degrees of intensity of the interactions by assigning a weight or strength to every edge. Such networks are called *weighted networks*. For the sake of simplicity, in this thesis I focus on undirected and unweighted simple graphs. However, all the results can easily be generalized to all other types of networks.

The simplest way to represent a network structure mathematically is by means of the adjacency matrix $\mathbb{A}$. This matrix is a square matrix of dimension equal to the number of nodes, N, in which every component, A_{ij}, encodes the interaction between nodes i and j. In my case of interest, simple unweighted undirected networks, all components are either 1, if they are connected, or 0, if they are not; and $\mathbb{A}$ is symmetric.

The process of converting a real system into its simple network representation involves its own difficulties and there is much interesting literature on the subject; but it is beyond the scope of this thesis. Here, I work on network datasets that are already constructed and do not discuss the nature of the nodes or their interaction in my analysis. Therefore, I do not focus on any particular real-world network, but present a set of theoretical studies that apply to any particular interconnected system. I only use real network datasets to assess whether my models and theories fit the types of structures that we can find in nature. However, I want to make it clear that it is important to always have the accurate definition of node and edge in mind when trying to apply the results obtained from network theory to a particular real system. Appendix A.1 gives some details of the empirical networks that I use in this thesis.

1.2.2 Degree distribution

In graph theory, the degree k of a node is the number of connections or neighbours it has, which in terms of the adjacency matrix can be calculated as: $k_i = \sum_j A_{ij}$. So, if a network has a number of nodes equal to N, and N_k of them have degree k, the probability that a randomly chosen node has degree k is given by:

$$P(k) = \frac{N_k}{N}. \tag{1.1}$$

The function $P(k)$ is called the degree distribution and it is the most fundamental statistical characteristic of a network. Degree distributions measured in real-world networks contrast with those expected from regular lattices or random networks. Typically, real networks have a scale-free degree distribution which exhibits a power law behaviour $P(k) \sim k^{-\gamma}$ with the exponent γ between 2 and 3 [45]. This type of distribution has a very fat tail and a second moment, $\langle k^2 \rangle$, that diverges in the thermodynamic limit. In a finite system, this fact leads to a standard deviation that is much larger than the mean. This heterogeneity implies that nodes that are at the tail of the distribution have many more connections than average: they are the hubs. These few, but significant, very well-connected nodes play a key role in network structure and function, and drive the behaviour of the whole system.

1.2.3 Degree correlations

Degree correlation among connected nodes is another important characteristic of network topology. Degree correlations are encoded in the joint distribution, $P(k,k')$, which gives the probability that an edge chosen at random connects two nodes of degree k and k'. Alternatively, one can use the conditional probability that an edge from a node of degree k connects it to a node of degree k': $P(k'|k)$. These functions are related by the expression:

$$P(k'|k) = \frac{P(k',k)}{\sum_{k'} P(k',k)} = \langle k \rangle \frac{P(k',k)}{kP(k)}, \tag{1.2}$$

where we use the fact that the fraction of links emanating from nodes of degree k is equal to: $\sum_{k'} P(k',k) = kP(k)/\langle k \rangle$. Moreover, the symmetry $P(k',k) = P(k,k')$ together with Eq. 1.2 leads to the detailed balance equation:

$$kP(k)P(k'|k) = k'P(k')P(k|k). \tag{1.3}$$

Both $P(k,k')$ and $P(k'|k)$ can be measured in real networks. However, in order to avoid strong fluctuations, there is a much less informative but convenient

measure: the average nearest neighbour degree of nodes of the same degree. This is given by:

$$\bar{k}_{nn}(k) = \sum_{k'} k' P(k'|k). \tag{1.4}$$

The function $\bar{k}_{nn}(k)$ relates the degree of a node to the average degree of its neighbours.

In the absence of correlations, the degree of the node at one end of an edge does not depend on the degree of the node at the other end. In this situation, the conditional probability has the form:

$$P(k'|k) = \frac{k' P(k')}{\langle k \rangle}, \tag{1.5}$$

and the average nearest neighbour degree becomes:

$$\bar{k}_{nn}(k) = \frac{\langle k^2 \rangle}{\langle k \rangle}, \tag{1.6}$$

which does not depend on the degree. If $\bar{k}_{nn}(k)$ depends on the degree, then there are correlations. In this situation, there are two possible cases. On the one hand, high-degree nodes can have a propensity to be connected to other high-degree nodes, the so-called assortative mixing, and $\bar{k}_{nn}(k)$ is an increasing function of the degree. On the other hand, high-degree nodes may preferentially to be connected to low-degree nodes, leading to disassortative mixing, so $\bar{k}_{nn}(k)$ decreases with the degree [22].

Real-world networks do have degree correlations. For example, it is well known that popular people also have popular friends. Measures determined on empirical datasets from social networks show assortative mixing by degree in practically all networks, corroborating this perception [52]. In contrast, technological and biological networks are more prone to disassortative mixing [116].

1.2.4 Clustering

Once correlations among pairs of nodes are described, we can take a step further and look at the three-point correlations. Correlations among three nodes translate into the tendency of nodes to have many common neighbours, creating triangles. This property is called clustering, and in graph theory is quantified by the clustering coefficient, c, defined as the probability that two randomly chosen neighbours of a node are connected. Given a node i with k_i neighbours, this probability is simply the number of edges between the neighbours of node i, t_i, divided by the total number of possible pairs of neighbours, $\frac{1}{2}k_i(k_i-1)$, so:

$$c_i = \frac{2t_i}{k_i(k_i-1)}. \tag{1.7}$$

If we average this coefficient over all the nodes of the same degree, we get the clustering spectrum:

$$\bar{c}(k) = \frac{1}{NP(k)} \sum_{i|k_i=k} c_i. \tag{1.8}$$

If we average over all the nodes instead, we obtain the local clustering coefficient:

$$\bar{c} = \frac{1}{N} \sum_i c_i = \sum_k P(k)\bar{c}(k). \tag{1.9}$$

An alternative to the local clustering coefficient is the global clustering coefficient, which measures the number of triangles present in the network over the total number of connected triples [135]; where a connected triple consists of three nodes connected by any path. The two definitions of the clustering coefficient are not equivalent and can give substantially different values. The local clustering coefficient measures the probability that two neighbours of a randomly chosen node are connected. The global clustering coefficient measures the density of triangles. Both measures are commonly used and in some situations can lead to opposite conclusions as to whether a network has a high or low degree of clustering.

These discrepancies emerge in networks with highly skewed degree distributions, similar to those present in real networks. On the one hand, the local clustering coefficient defined in Eq. 1.9 tends to be dominated by nodes of low degree, which usually have a higher value of the clustering coefficient [150]. On the other hand, the number of connected triples diverges in the thermodynamic limit for networks with a power law degree distribution with an exponent $2 < \gamma < 3$, leading to very small values of the global clustering coefficient [62]. In these circumstances, I prefer not to focus on a single clustering coefficient of the whole network, but rather I consider the clustering spectrum. Analysing clustering for all nodes of the same degree does away with the discrepancies. Along these lines, in this thesis I usually consider the whole clustering spectrum; and when I refer to the clustering coefficient of a network, I am referring to the local clustering coefficient.

This tendency of nodes to have many common neighbours is clearly one of the main features of social networks [131]. Moreover, high levels of clustering have been repeatedly found in networks with completely different natures [123]. Figure 1.3 shows the clustering spectra of three real paradigmatic networks from different domains: infrastructure, social, technological, and biological. If we compare the clustering spectra of the networks with a randomized version of them, with the same degree sequence, we can see that the real networks have higher degrees of clustering than those expected by chance.

There are other important topological features present in many real networks, such as the small world property [168, 170], or the community struc-

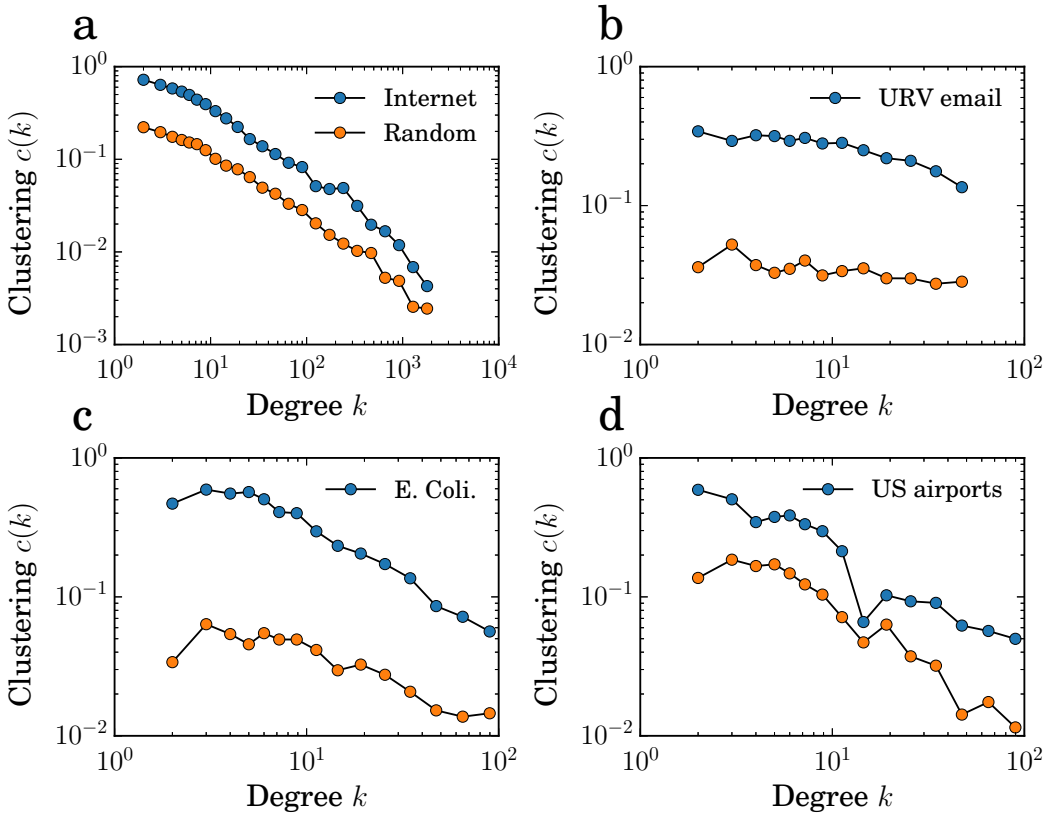


Figure 1.3: Clustering spectra of four real paradigmatic networks from different domains (blue) compared with their randomized versions with the same degree sequence (orange). **a)** The autonomous system (AS) Internet topology for June 2009. **b)** The social email network of the Rovira i Virgili University. **c)** The metabolic network of the bacterium E. Coli. **d)** The US air transportation network. See Appendix A.1 for more details of the networks. The method used to randomize the networks is explained in section 2.6.

ture [56, 133, 136]. Nonetheless, the main goal of this thesis is to discuss the origin and effects of clustering on the structure and function of networks, so those other features are beyond the scope of this work.

1.3 Network models

The description and measurement of real-world network structures has revealed some very important common features. Indeed, the small world property, broad degree distributions, degree correlations and high levels of clustering are all repeatedly found in networks of completely different origins. However, in order to understand the emergence of these features and the effect they have on network structure and function, we need to build network models.

We can use static models that allow us to create networks with some particular properties of interest and that are random in all other respects. Such models are very useful to separate a specific topological feature from the others in order to study its effect on them and also on network processes. However, networks are not static but evolve: they change over time. This evolution of networks is a product of dynamical processes that create and remove both nodes and edges, and which typically is unplanned and decentralized. Mechanisms of network formation normally depend on interactions between agents that operate at a local scale giving rise to a self-organized system. This fact encouraged scientist to develop growing network models that, based on simple rules, can reproduce the evolution of real-world network structures.

Moreover, the evolution of networks is not deterministic but stochastic. For instance, if we could go back to 1969 and repeat the process of formation of the physical Internet network, with the same initial condition, we would not observe exactly the same topology. We reckon that, although there would be important common large-scale features, not all the edges would connect exactly the same pairs of nodes. The Internet that we observe today is just one possible outcome: represented by one of a set of possible graphs, or ensemble of graphs, where each graph has its own given probability of existence. Unfortunately, we do not have access to the whole ensemble of graphs that real networks could be represented by, but just one instance. Thus, it is impossible to explore the patterns that drive network formation only from observational data. Therefore, we need to develop network models that reproduce the properties of real-world networks and assume that the mechanisms that define the models are similar to those that generated the real networks.

Much effort has gone into developing static and growing network models; and the definition and characterization of random models have constituted one of the main areas of research in network science.

1.3.1 Classical random graph

The first and simplest example of a random graph is the *Erdős-Rény model* which consists of the ensemble of networks $G_{N,E}$, whose members have the same number of nodes, N, and the same number of edges, E [67, 66]. To generate a random network from this ensemble we have to spread our E edges among the $N(N-1)/2$ pairs of nodes at random. Some properties of this ensemble are straightforward to calculate, e.g., the average degree $\langle k \rangle = \frac{2E}{N}$. Unfortunately, other properties that we are interested in, such as the degree distribution or the clustering coefficient, are not so easy to find analytically.

Fortunately, physicist know that analysis of the canonical ensemble is technically easier than analysis of the micro-canonical ensemble. Hence, if we consider that edges play the role of energy, the Erdős-Rényi model corresponds to the micro-canonical ensemble once the number of edges is fixed. In this scheme, the canonical ensemble would be a model in which, instead of fixing the exact number of edges, we fixed the average number of them. This would involve fixing a probability of connection between all the pairs of nodes. This ensemble of networks, $G_{N,p}$, although it is known as the Gilbert model, was first introduced by Solomonoff and Rapoport [162, 163] and is defined by just two parameters: the number of nodes N, and the connection probability p. So to generate a network from this ensemble we have to visit the $N(N-1)/2$ pairs of nodes and connect them with the fixed probability p. The average number of edges in this ensemble is directly the fraction, p, of pairs that we actually connect:

$$\langle E \rangle = p\frac{N(N-1)}{2}. \tag{1.10}$$

Using the grand canonical ensemble we can now calculate the probability that a node has k connections, which comes directly from the binomial distribution:

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{1.11}$$

The average degree is then $p(N-1)$, which is the fraction, p, of all the $N-1$ nodes that a node can be connected to. When the network is large ($N \to \infty$), while $\langle k \rangle$ is fixed, the binomial distribution approaches the Poisson distribution:

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \tag{1.12}$$

This grand canonical version of the random graph also allows us to calculate the clustering coefficient. We first need to recall that the clustering coefficient is the probability that two neighbours of a node are themselves neighbours too.

Because all pairs of nodes are connected independently from the rest, the clustering coefficient is equal to p:

$$\bar{c} = p = \frac{\langle k \rangle}{N-1}. \tag{1.13}$$

So the clustering coefficient for an infinite sparse classical random graph approaches zero. That is, clustering in these networks is only a finite-size effect.

Both the fast decay of the degree distribution and the vanishing level of clustering generated by the classical random graph are in contrast with the common occurrence in empirical data of scale-free degree distributions and high levels of clustering. This fact implies that the connectivity patterns of real networks, although they are not as regular as lattices, are not completely random. Therefore, although the classical random model is very simple, it already reveals that there are some underlying mechanisms of formation that are the origin of the particular topological properties of real networks.

1.3.2 Preferential attachment

In the 1970s, Derek de Solla Price addressed the problem of the heavy-tailed degree distributions of real-world networks not reproduced by the classical random graph. He presented a novel model based on a very simple idea: a growing model in which new nodes appear and connect to a fix number of already existing nodes proportionally to the number of connections of those already existing nodes [59]. This formation pattern introduces a "rich-get-richer" effect which increases the number of hubs and leads to a much broader degree distribution than that of classical random graphs. In 1999, Barabási and Albert published a similar model in which the mechanism was called *preferential attachment*, which has become the accepted name in recent years [9]. In this model, at each time step, a node appears and creates m new connections with already existing nodes with a probability proportional to the current degree of those already existing nodes. In this case, when the network becomes large ($t \to \infty$), the degree distribution can be calculated easily. The result is a stationary power law distribution, $P(k) \sim k^{-\gamma}$, with exponent $\gamma = 3$: a scale-free distribution similar to those observed in many real networks. Of course this particular exponent is obtained with simple proportionality of the preferential attachment; to achieve arbitrary exponents, we simply substitute that proportionality for a linear function [107].

Subsequent work, both empirical and analytic, studied other topological properties of the Barabasi-Albert model. For example [77, 106] show that the clustering coefficient of these network models vanishes in the thermodynamic limit, in contrast to the high level of clustering present in empirical networks.

However, the simplest form of the preferential attachment model already reveals that the scale-free degree distributions observed in real-world networks

are just a manifestation of a certain kind of rich-get-richer mechanism in the formation of their connections.

1.3.3 The configuration model

Once the underlying mechanism of the broad degree distribution is revealed, we can move forward and study other random network models in which the degree distribution is a constraint and which are random to all other respects. Along these lines, by the late 1970s, network scientists developed what is now known as the *configuration model* [17, 31]. In this model, a certain number of half edges, or stubs, are distributed among the nodes, according to a given degree sequence. Then, pairs of stubs are connected at random, creating edges, until there are no stubs left. This procedure corresponds to building a maximally random network with a given degree sequence.

In principle, in this model every possible matching should be given exactly the same probability. However, during the network formation process a key issue arises: how do we deal with the possibility of creating more than one edge between nodes, multi-edges, or edges that connect nodes to themselves, self-edges? On the one hand, if we ignore this issue, we do not obtain exactly the target degree sequence. On the other hand, rejection of such edges introduces a bias in the configuration space that leads to non-uniform sampling from the ensemble. Some prefer to ignore this issue, considering that the number of these types of edges vanishes in the thermodynamic limit. However, multi-edges are concentrated between nodes with a high degree, and in the case of very heterogeneous networks, they do have an important effect [26]. The restriction of the tendency of high-degree nodes to have multi-edges creates connections between high-degree and low-degree nodes instead, inducing negative degree correlations [142]. Nonetheless, this disassortativity is observed in many technological and biological networks [116]. Hence, part of the observed tendency of hubs to be connected to low-degree nodes can arise from a topological constraint rather than a specific formation mechanism. In contrast, the positive degree correlation measured in social networks suggests that a specific pattern formation mechanisms is responsible for the propensity of popular people to be connected to other popular people.

The density of triangles generated by the configuration model can be calculated analytically. Precisely, the local clustering coefficient is given by the expression:

$$\bar{c} = \frac{\langle k(k-1)\rangle^2}{N\langle k\rangle^3}, \tag{1.14}$$

and therefore vanishes for large systems [21, 135]. However, Eq. 1.14 does not take into account the existence of multi-edges and self-edges. Thus, for highly

heterogeneous degree distributions, Eq. 1.14 becomes incorrect, leading to a non-physical solution ($\bar{c} > 1$). Unfortunately, the model in which multi-edges and self-edges are forbidden does not admit an analytic solution. This problem can be addressed by using the canonical version of the configuration model in which it is not the exact degree of each node that is fixed, but the average degree. In chapter 3, I use this scenario to find the correct analytic expression for the clustering coefficient and, together with numerical simulations, show that in some particular cases the configuration model can give larger densities of triangles than is commonly believed possible.

Notwithstanding these shortcomings, the configuration model has been of vital importance because it isolates the degree distribution from the other topological properties. Therefore, the configuration model is the best framework for studying the effect of the degree distribution on other topological properties and on network process. The most important contribution that has arisen from the study of the configuration model is the discovery of the vanishing epidemic and percolation threshold of scale-free networks with a degree exponent $\gamma \leq 3$ [37, 46, 144]. This implies that, due to the heterogeneity of the number of connections of nodes, real networks are highly robust against random failure of their constituents, but at the same time they can propagate any infectious agents.

1.3.4 Clustered network models

None of the random network models that I have introduced so far generates a level of clustering comparable to that observed in real networks. Much effort has been devoted to developing models of clustered networks like the Watts Strogatz model [170], geometric network models [108, 156] or a set of random modular graph models [84, 98, 134]. Among them, geometric models have contributed the most to our understanding of the nature of clustering.

In general terms, geometric network models assume that all network nodes reside in an underlying hidden metric space in which the distance between two nodes represents a cost for a connection to exist. This corresponds to assuming that similarity among nodes is encoded in this metric space, so connections are more probable between closer, or more similar, nodes. In this framework, clustering arises as a natural consequence of the triangle inequality in the underlying geometry. Moreover, if the underlying space is hyperbolic, instead of Euclidean, the resulting networks are also small-world networks and have a scale-free degree distribution [108, 156]. Therefore, the existence of a metric space is the best candidate formation mechanism behind the high concentration of triangles in real networks.

However, the geometric scheme is not suited to the study of processes and

the dynamics of networks. The other models are not very good at reproducing properties of real networks other than clustering, so they are not appropriate for the study of the effect of clustering on the structure and function of complex networks. Therefore, the development of new models of generation of clustered networks is one of the major challenges that network science faces.

1.4 Processes on networks

The final goal of the study of the structure of complex networks is to understand their function and how they would behave under any kind of process that may occur within them. For example, we study the structure of social networks to understand better how diseases and rumours spread over those networks, in order to design better vaccinations or marketing strategies. We study the connectivity patterns of the Internet to design better protocols for routing information through the network. We are interested in the topology of power grids in order to be able to construct more robust networks and thereby avoid blackouts. We study neural networks in order to discover the conditions under which neurons globally synchronize and trigger an epilepsy attack.

Much research has attempted to make the connection between the structure and function of networks, and there has been valuable progress in some areas. Many theories and models have been developed that describe processes and at the same time represent the role that connectivity patters play in network phenomena. Of all the possible processes, the failure of constituents and epidemic spreading are the fields of network science that have have received most attention. Here, I focus on one of the simplest network processes: percolation. It provides an elegant theory of the robustness of networks to the failure of their constituents and at the same time it is directly related to the spread of infectious agents.

1.4.1 Percolation

The main concern in the design of technology and infrastructures is robustness. A robust network requires many alternative paths among nodes, in order to maintain global connectivity even if some nodes or connections stop functioning for some reason. For instance, within the Internet, a certain proportion of routers are not functioning at any time and yet data packages are successfully rerouted and delivered to their destinations correctly. However, there are concerns among Internet experts that the existing Internet routing architecture may become unsuitable to meet the demands placed on within a period as short as the next decade [120].

In order to study the robustness of a network, one has to measure how the failure of nodes and edges affects the largest number of nodes that are connected by any path (the largest connected component[1]). If these failures are due to random breakdowns or attacks, then this phenomenon can be modelled by a percolation process.

Percolation is a classic problem that has attracted the attention of mathematicians and physicists for many years because it is one of the simplest models that displays a phase transition. The percolation problem on a network can be stated as follows: we visit each node (site percolation) or each edge (bond percolation); with a probability of p we preserve it and with a probability of $1-p$ we remove it. Under this process, complex networks undergo a continuous phase transition at a critical value, p_c , known as the percolation threshold. Below p_c, the network is made of a myriad of finite disconnected clusters. Above this critical value, a macroscopic cluster of the order of the size of the system emerges, so the network becomes globally connected. In practical applications, an accurate prediction of the percolation threshold is extremely important. In the case of infrastructure or technology networks, this threshold defines the robustness of the network to random failures of its constituents. Moreover, in epidemiology, it represents whether a disease will die out or reach an endemic steady state.

In some cases, such as the 1-dimensional lattice, the 2-dimensional square lattice or the Bethe lattice, the percolation threshold has been found analytically [165]. In the case of random networks, one can use the absence of short loops to find an approximate expression for the percolation threshold in terms of some topological properties. However, real network do have a high density of short loops, so the theories that have been developed still do not yield very accurate predictions on real cases [71]. For example, Fig. 1.4 compares bond percolation simulations of the western states of the United States power grid with the most accurate theory developed so far, in 2014 [99]. As we can clearly see, the theoretical curve deviates greatly from the numerical simulations. Therefore, the next step is to include clustering in new theoretical approaches and develop new clustered network models that enable us to study exactly how the presence of triangles affects the bond percolation properties of networks.

[1] In this thesis we refer as a component or a cluster to a subgraph in which every pair of nodes are connected by at least one path. If a network has two clusters, or components, means that there are two parts of the networks that are completely disconnected; so there is no path from a node of one part to any node of the other part.

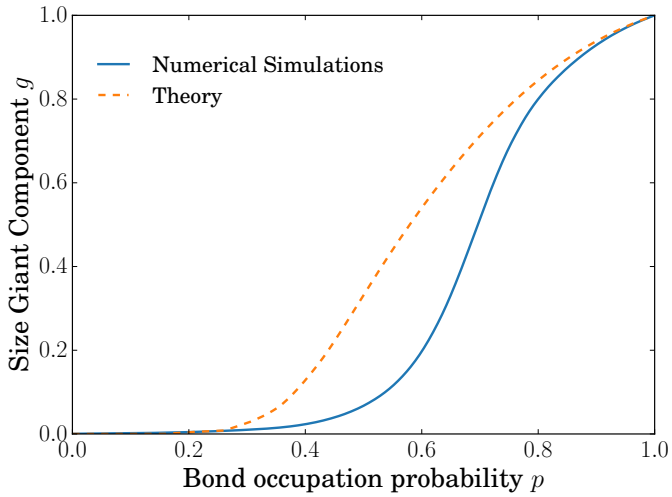


Figure 1.4: Bond percolation simulation for the western states of the US power grid network compared with the theoretical curve developed in [99]. Relative size of the largest connected component as a function of the bond occupation probability, p.

1.5 Outline of the thesis

The random graph models that I have introduced were important breakthroughs in network science because they shed light on both the origin of the degree distribution and its effects on other topological properties and network processes. Along the same lines, much other work introduced modifications to those models to study, both analytically and empirically, the role of the degree correlations of networks in the topology and dynamics of networks [20, 24, 78, 142, 169]. However, the high level of clustering present in empirical networks still remains one of their least understood features.

Study of the hidden metric spaces that underlie complex networks has represented an important step forward in our understanding of the mechanisms behind the origin of clustering [27, 28, 158]. Nonetheless, little is known about the effect that the presence of short loops can have on other topological properties, or how it affects epidemic spreading and the robustness of networks. The reason for this is twofold. First, the mere presence of triangles in a network contradicts the assumption that networks are locally tree-like. That assumption is used almost across the board in mathematical tools applied in network theory. Therefore, the violation of the assumption hinders any possible proper theoretical framework for clustered networks. Second, there is a lack of realistic clustered network models that allow the study of the effect of clustering on the

structure and function of complex networks.

The most immediate example of the need for new insight into the role clustering plays in the properties of networks is the classic and simple, though crucial, bond percolation problem. When we compare current state-of-the-art theories with numerical simulations on real networks, it becomes obvious that we are still a long way from a precise description of real phenomenon. Since available theories are highly accurate for locally tree-like networks, it seems obvious that clustering is a vital missing piece of the puzzle [71]. Much work addresses this issue, but the difficulty in resolving it becomes clear once we see the large diversity of opinions [84, 85, 86, 100, 104, 117, 122, 132, 134, 153, 154, 155].

Given this framework, this work aims to contribute to our understanding of the role played by clustering in the structure and function of complex networks. In order to do that, in chapter 2, I start by introducing a class of network models, exponential random graphs, which will form the fundamental framework of my study. One of the most important applications of exponential random graphs is the use of maximally random graphs with an expected degree sequence. I discuss this model and later, in chapter 3, I use it to analyse the clustering generated in random scale-free networks.

I will also use exponential random graphs to present a new network model that is capable of representing high levels of clustering and in which triangles are organized in the most random way. After that, in chapter 4, I compare my clustered network model with the already existing models and analyse which of them best reproduce real-world networks.

Then in chapter 6, I use my model to study how clustering affects the position of the bond percolation threshold. I reveal that clustering can have a striking effect on the structure of networks at the macroscopic level that completely redefines their percolation phase space. More precisely, I show that clustered networks can undergo not one, but many percolation transitions; this reveals a new phenomenon in complex networks.

Finally, in chapter 7, I develop a new method to find real networks that undergo multiple percolation phase transitions. This reinterpretation of the most recent percolation theories offers new insight into the accuracy of the theoretical values of bond percolation thresholds.

This work contributes to our understanding of the role clustering plays in network structure and function. In particular, my maximally random clustered network model is an important breakthrough in network science, since it defines an appropriate framework within which to study the effect that clustering has on network topology and network processes. Furthermore, the new multiple percolation transition phenomenon that I reveal completely changes how we should tackle the percolation problem on complex networks, with major implications for network robustness and epidemic spreading.

CHAPTER 2

Exponential random graphs

2.1 Ensembles of networks and exponential random graphs

Network models are generally defined by a network generator mechanism. These mechanisms are typically stochastic, so a network model can also be defined by what is know as a network ensemble, which is a set of possible graphs $\{G\}$ in which each graph has a probability $P(G)$ to be found in a particular realization. Once we have a well defined ensemble we can calculate the expected value of any topological property of a network sampled from this ensemble, x, simply by $\langle x \rangle = \sum_{\{G\}} P(G)x(G)$. Such observables can be any measurable topological property of a network, e.g. the number of edges, degree sequence, clustering coefficient, average path length, etc... Here, we assume that the observables of a network that belongs to the ensemble are $\{x_i\}$ and their values measured in the considered real network are $\{x_i^*\}$. In our case we are just considering simple graphs with a certain number of nodes N.

As an example, in the classical random graph, the constrains are just the number of edges E, so $\{x_i^*\} = E$. In the Erdős-Rényi ensemble, $G_{N,E}$, there are only graphs that have N nodes and E edges and all these graphs have exactly the same probability to exist. Therefore, in this case, the ensemble contains all the possible networks with exactly the same observables of the real network, so $\{x_i\}(G) = E \ \forall G \in \{G\}$, and $P(G)$ is a constant independent of G. Instead, in the Gilbert model, we fix the number of nodes and a probability p of connection between nodes. Hence, in this ensemble $G_{N,p}$, there are all the possible graphs with N nodes, and each graph have the probability to exist equal to $P(G) = \prod_{i<j} p^{A_{ij}}(1-p)^{1-A_{ij}}$, where A_{ij} is a component of the adjacency matrix $\mathbb{A}$ of the graph G. In such situation, the network ensemble does not fix the exact number of edges but its expected value, $E = x_i^* = \langle x_i \rangle = pN(N-1)/2$. In the case of the configuration model, the constrains are now the degree of all nodes, so $\{x_i^*\} = \{k_1, k_2, ..., k_N\}$. Thus, we give to each node a certain number of stubs according to the degree sequence $\{x_i^*\}$. Therefore, the configuration model ensemble only contains the graphs that agree with this degree sequence, so $\{x_i\}(G) = \{x_i^*\} \ \forall G \in \{G\}$, and all have the same probability to exist.

These examples make clear the difference between the micro-canonical and

the canonical ensembles. On the one hand, in a micro-canonical ensemble the set of graphs is restricted to those with observables equal to the constrains, $\{x_i\}(G) = \{x_i^*\}\ \forall G$, and the probabilities of realizations, $P(G)$, are the same for all of them. On the other hand, a canonical ensemble contains all possible graphs [1] and the probability of each graph to exist is defined in such a way that the expected value of their observables are equal to the constrains, $\langle x_i \rangle = x_i^*$. Although the canonical ensemble looks less accurate than the micro-canonical ensemble, in the thermodynamic limit, and for sparse networks, both ensembles are equivalent [6]. Besides, the canonical ensemble allows for analytic treatment in some situations that is not possible in the micro-canonical ensemble. Thus, canonical ensembles are more convenient for theoretical analysis.

Network models are useful for two different purposes. On the one hand, some network models can reproduce real networks observables beyond the ones that are fixed by the constrains, $\{x_i^*\}$. This type of models are normally growing models, and we expect that the formation mechanism that defines the model is of the same nature that the one that drives the real network evolution. This is the case of the Barabási-Albert model and the preferential attachment mechanism, in which just fixing the number of nodes and the number of links it is able to reproduce the scale-free behaviour of real networks. On the other hand, there are other models that are not giving insights into the network formation but are able to generate networks in a very controlled way. This types of models are very useful when studying the effect of one particular topological property on processes that occur on top of networks. This is the case of the configuration model that has given many insights into the effect of the degree distribution on network robustness and networks processes, like the spread of diseases.

In the second scenario, there are the ensembles of networks that we introduce here, the exponential random graph models (ERG). ERG are canonical ensembles of networks characterized by the probability of each graph to exist that have the form $P(G) \propto e^{H(G)}$. The function $H(G)$ is the Hamiltonian of each graph G and is given by $H(G) = \sum_i \alpha_i x_i(G)$, where $\{x_i\}$ are the observables of the graph G and $\{\alpha_i\}$ are Lagrange multipliers such that the average value of the observables are equal to the constrains, $x_i^* = \langle x_i \rangle = \sum_{\{G\}} P(G) x_i(G)$. Here we are going to explain in detail some applications of ERG that are going to be frequently used throughout this thesis.

[1] Note that when we refer to all possible graph we are always restricted to simple graphs with number of nodes equal to N.

2.2 Maximally random graphs

Once we define a model with given constrains, we expect that all other topological features are left to randomness. In other words, we want maximally random graphs that fulfils a certain set of constrains. Therefore, due to maximum entropy principle of information theory [53, 160], and from the second law of thermodynamics in statistical physics [147], the best choice of the probability distribution $P(G)$ is the one that maximizes the Shannon/Gibbs entropy,

$$S = -\sum_{\{G\}} P(G) \ln P(G). \tag{2.1}$$

As a canonical ensemble, the observables $\{x_i\}$ of the networks should be equal in average to the observables of the real network we are trying to model. So

$$x_i^* = \langle x_i \rangle = \sum_{\{G\}} P(G) x_i(G) \quad \forall i. \tag{2.2}$$

At the same time $P(G)$ must be normalized

$$\sum_{\{G\}} P(G) = 1. \tag{2.3}$$

We can introduce these constrains using Lagrange multipliers when maximizing the entropy

$$\frac{\partial}{\partial P(G)} \left[S - \gamma \left(1 - \sum_{\{G\}} P(G) \right) - \sum_i \alpha_i \left(x_i^* - \sum_{\{G\}} P(G) x_i(G) \right) \right] = 0 \quad \forall G \in \{G\}. \tag{2.4}$$

After the derivative we obtain

$$-\ln P(G) - 1 + \gamma + \sum_i \alpha_i x_i(G) = 0, \tag{2.5}$$

which implies

$$P(G) = e^{-1+\gamma+\sum_i \alpha_i x_i(G)} = \frac{e^{\sum_i \alpha_i x_i(G)}}{e^{1-\gamma}} = \frac{e^{H(G)}}{Z} \tag{2.6}$$

which have the same form of the ERG so $H(G) = \sum_i \alpha_i x_i(G)$ is the structural Hamiltonian of the network and $Z = e^{1-\gamma}$ is the partition function [3, 4, 80]. Equation 2.6 shows that the exponential random graphs that we introduced are graphs that fulfil a given set of constrains $\{x_i^*\}$ and are maximally random to all other respects.

2.3 Maximally random graphs with expected degree sequence

As a first example of an application of the exponential random graph models (ERG) we consider the canonical version of the configuration model. This case of interest corresponds to fix the expected degree of each node, so the observables that the ensemble has to reproduce are $\{x_i^*\} = \{k_1, k_2, ..., k_N\}$. Note that, because the ensemble is canonical, we are fixing the observables on average. Then, if we express the degree of a node as a sum, using the adjacency matrix as $k_i = \sum_j A_{ij}$, from Eq. 2.6 we can calculate the probability $P(G)$ to find a particular graph G with a given adjacency matrix $\mathbb{A}$ as

$$P(G) = \frac{e^{\sum_i \alpha_i \sum_j A_{ij}}}{Z} = \prod_{i<j} \frac{e^{(\alpha_i+\alpha_j)A_{ij}}}{1+e^{(\alpha_i+\alpha_j)}}. \tag{2.7}$$

Notice that this expression can be re-written as

$$P(G) = \prod_{i<j} r_{ij}^{A_{ij}} (1-r_{ij})^{1-A_{ij}}, \tag{2.8}$$

where

$$r_{ij} = \frac{e^{(\alpha_i+\alpha_j)}}{1+e^{(\alpha_i+\alpha_j)}} \tag{2.9}$$

is the probability of the existence of a link between nodes i and j. Notice that the factorization in Eq. (2.8) implies that a network belonging to this ensemble can be generated by pairwise connection probabilities given by Eq. (2.9). Finally, by redefining the Lagrange multipliers as $\kappa_i = \kappa_s e^{\alpha_i}$, we obtain the following connection probability

$$r\left(\frac{\kappa\kappa'}{\kappa_s^2}\right) = \frac{\kappa\kappa'}{\kappa_s^2}\left(1+\frac{\kappa\kappa'}{\kappa_s^2}\right)^{-1}, \tag{2.10}$$

where κ and κ' are the transformed Lagrange multipliers associated to each node, and κ_s is a parameter that its the same for all nodes and we will give an interpretation afterwards.

To generate networks from this ensemble we visit each pair of nodes and connect them with the probability given by Eq. 2.10. However, we must first find the relation between the Lagrange multipliers, or the hidden variable κ [22], and the expected degree of a node, that is, the constrain of our model.

Given the connection probability of Eq. 2.10, a node i of hidden variable κ_i will have an average number of connections $\bar{k}(\kappa_i)$ equal to

$$\bar{k}(\kappa_i) = \sum_{j=0}^{N} \langle A_{ij} \rangle = \sum_{j=0}^{N} r_{ij} = \sum_{j=0}^{N} \frac{\kappa_i\kappa_j}{\kappa_s^2}\left(1+\frac{\kappa_i\kappa_j}{\kappa_s^2}\right)^{-1}. \tag{2.11}$$

Here, we use the ansatz $\kappa_i\kappa_j << \kappa_s^2$ so the probability of connections can be approximated by $r_{ij} \sim \kappa_i\kappa_j/\kappa_s^2$ and the average degree of a node becomes

$$\bar{k}(\kappa_i) \sim \sum_{j=0}^{N} \frac{\kappa_i\kappa_j}{\kappa_s^2} = \frac{\kappa_i}{\kappa_s^2}\sum_{j=0}^{N}\kappa_j. \tag{2.12}$$

So if we fix the constant $\kappa_s^2 = \sum_{j=0}^{N}\kappa_j$, which is the same for all nodes, then the expected degree of a node is equal to its transformed Lagrange multiplier, $\bar{k}(\kappa) = \kappa$ [22, 142, 159]. This is why the transformed Lagrange multiplier κ its normally called hidden degree [156].

Then, we can calculate the parameter κ_s which is

$$\kappa_s = \sqrt{\sum_{j=0}^{N}\kappa_j} = \sqrt{\sum_{j=0}^{N}\bar{k}_j} = \sqrt{2E} = \sqrt{N\langle k\rangle}. \tag{2.13}$$

So κ_s is directly related to the total number of edges of the network E. From Eq. 2.13 we can now revised the ansatz used to obtain Eq. 2.12 and conclude that the ansatz is strictly valid only if there is no node with expected degree larger than $\sqrt{N\langle k\rangle}$. This fact implies that the parameter κ_s is a structural cut-off defining the onset of structural correlations, that is, nodes with expected degrees below κ_s are connected with probability $r(\frac{\kappa\kappa'}{\kappa_s^2}) \sim \frac{\kappa\kappa'}{\kappa_s^2}$ and, therefore, are uncorrelated at the level of degrees. This structural cut-off is also present in the microcanonical version of the configuration model [39].

This structural cut-off, κ_s not only have important effects on the degree correlations but also on the higher order correlations like the clustering coefficient. As we will see in the next chapter, the finite size scaling of the clustering coefficient of the maximally random graphs with expected degree sequence strongly depends on the value of the maximum degree of the network k_c.

Now that we can calculate the Lagrange multiplier of each node, we can now sample networks from this ensemble. To do this we just have to fix the hidden degree of each node equal to its expected degree. Then, we visit each pair of nodes and connect them with probability r_{ij} given by the Eq. 2.10.

2.4 Monte Carlo sampling from exponential random graph ensembles

Sampling from the canonical ensemble of the configuration model that we have just described is quite simple. This is due to the fact that we were able to factorize the probability of sampling one graph, $P(G)$, as a product of the probability of links between nodes to exist. However, for many given constrains $\{x_i^*\}$ we can

not tackle the problem analytically, so we have to rely on numerical solutions. In these situations, we can take advantage of Monte Carlo simulations, which are suited for ERG.

The exponential form of $P(G)$ makes the Markov chain Monte Carlo (MCMC) method together with the Metropolis-Hastings algorithm the best way for sampling networks from a ERG ensemble [92, 119]. In this algorithm, one has to define a move in the graph space. Examples of this move, that is able to change one graph G to another graph G' of the ensemble, include removing, changing, or swapping edges. Then, starting from an initial graph, a move is proposed and it is always accepted if $H(G') > H(G)$ and with probability $P(G')/P(G)$ instead. So the probability to accept a move is

$$p = \min(1, \frac{P(G')}{P(G)}) = \min(1, e^{H(G')-H(G)}) = \min(1, e^{\Delta H}). \tag{2.14}$$

At the steady state a new graph is sampled from the ERG ensemble defined by the Hamiltonian H.

This method makes the sampling of maximally random networks with any desired observable very easy. Let's choose a certain observable O^* that a network of the ensemble should reproduce. Then, we define a Hamiltonian that depends on the difference between the target observable and the observable O of the current network

$$H = -\beta|O^* - O|. \tag{2.15}$$

This ensemble corresponds to an ERG ensemble in which networks have the observable O close to the target observable O^* as a function of the parameter β, which can be interpreted as an inverse of the temperature. If we are in a hot regime, the average observable of the networks is going to be far from our target observable, and in the cold regime we will sample networks with an observable very similar to the target one.

So to generate maximally random networks that reproduces a real world network property we take the empirical network as the initial graph and we apply the Metropolis-Hastings algorithm as defined above. However, the moves in the graph space are very local because involves only the modification of just few edges. Thus, we need to make use of a simulated annealing procedure in order to make sure that we are able to visit all the graph space [40, 103]. This implies that we will start sampling networks from the hot regime and we are going to decrease the temperature slowly until we are sampling networks with an expected value of the observable very close to the target one.

2.5 Maximally random clustered networks

As we have seen, we can use ERG together with Monte Carlo simulations to generate maximally random networks with any topological property similar to real networks. The aim of this thesis is to study the effects of clustering on other topological properties and on the bond percolation process. Therefore, the possibility to generate maximally random networks with a precise control on the level of clustering defines the best framework for such study.

The Monte Carlo step that we decided to use in order to move into the graph space is the rewiring of edges. We use two different rewiring schemes. In the first one, two different edges are chosen at random. Let these connect nodes A with B and C with D. Then, the two edges are swapped so that nodes A and D, on the one hand, and C and B, on the other, are now connected. We take care that no self-connections or multiple connections between the same pair of nodes are induced by this process. This rewiring scheme preserves the degree sequence of the original network but not degree-degree correlations. In the second rewiring scheme, we first chose an edge at random and look at the degree of one of its attached nodes, k. Then, a second link attached to a node of the same degree k is chosen and the two links are swapped as before. Notice that this procedure preserves both the degree of each node and the actual nodes' degrees at the end of the two original edges. Therefore, the procedure preserves the full degree-degree correlation structure encoded in the joint distribution $P(k,k')$. Both procedures are ergodic and satisfy detailed balance [55, 160].

Regardless of the rewiring scheme at use, the process is biased so that generated graphs belong to an exponential ensemble of graphs $G \in \{G\}$, where β is the inverse of the temperature and $H(G)$ is a Hamiltonian that depends on the current network configuration. In case we are interested in fixing the clustering coefficient we choose a Hamiltonian that depends on the target local clustering coefficient $\bar{c}$ as

$$H = -\beta|\bar{c}^* - \bar{c}| \tag{2.16}$$

where $\bar{c}^*$ is the target local clustering clustering coefficient and $\bar{c}$ the current one. However, if we are not interested on the clustering coefficient but in the clustering spectrum then the Hamiltonian takes the form

$$H = -\beta \sum_{k=k_{min}}^{k_c} |\bar{c}^*(k) - \bar{c}(k)|, \tag{2.17}$$

where $\bar{c}^*(k)$ is target degree-dependent clustering coefficient and $\bar{c}(k)$ the current one. We then use the simulated annealing algorithm based on a standard Metropolis-Hastings procedure. We first start by rewiring the network $200E$ times at $\beta = 0$, where E is the total number of edges of the network. Then, we

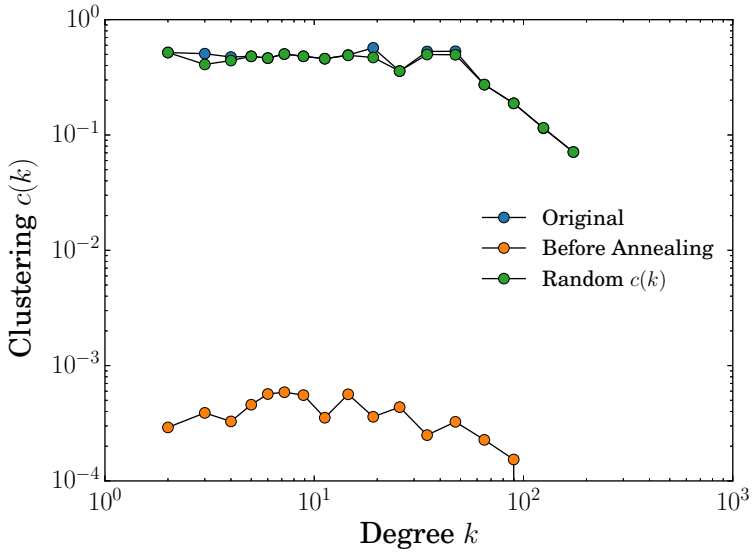


Figure 2.1: Comparison of the clustering spectrum of the PGP social Network (blue), its maximally random version with same degree distribution and clustering spectrum (green), and the starting network of the annealing process after $200E$ rewires at $\beta = 0$ (orange).

start an annealing procedure at $\beta_0 = 50$, increasing the parameter β by a 10% after each $200E$ rewiring events. We keep increasing β until the target clustering spectrum is reached within a predefined precision or no further improvement can be achieved.

As an example, Fig. 2.1 shows the comparison of the levels of clustering of the PGP social network and its maximally random network version where we fixed the clustering spectrum while preserving the degree sequence during the rewiring procedure (See the Appendix A.1.2 for a description of the network data set).

As we can clearly see our maximally random clustered network model reproduces with a very good accuracy the target clustering spectrum. This good performance is much better than previous similar models [82]. Besides, the low level of clustering of the network before starting the annealed process vanish any possible concern of a lack of ergodicity in our model.

There are some concerns that sampling networks using the rewiring method fixing the degree sequence or degree correlations is not completely uniform [6, 51, 151]. In these works they propose a modification to guarantee the exact uniform sampling. However, we do not use these modification because our edge swapping sampling is close to uniform in the graphs that we study in this thesis, and the effects in real cases between the two types of dynamics are seen to

be negligible [6, 124]. Moreover, these modifications have a computational cost that in some situation are not affordable. Nonetheless, we realize that an efficient implementation of these modifications in order to guarantee the uniformity in the graph sampling of our model could be an interesting improvement that will be studied in the future.

2.6 RandGenNet

In order to promote the used of the ERGs we developed the program RandNetGen that is able to generate maximally random networks with any desired topological property. This program randomizes an initial undirected and unweighted network and, using the biased rewiring mechanism described in the previous sections, can fix some topological properties. These network properties to fix can be similar to those of the original network or given as a file so they can have any desired value. The list of possible network properties that the program can fix are:

- The original joint degree distribution $P(k,k')$
- The average neighbour degree $K_{nn}(k)$
- The local clustering coefficient $\bar{c}$
- The number of triangles (global clustering coefficient)
- The clustering spectrum $\bar{c}(k)$

The code is available on the collaborative code web page Github:
(http://polcolomer.github.io/RandNetGen).

The possibility of fixing a network property to any desired value makes RandNetGen a very useful tool for studying empirically the effect that a certain topological property has on other topological properties and network processes. Specifically, in this thesis we used RandNetGen to study the effect of clustering on the percolation process by generating networks with the same degree distribution and degree correlations but different levels of clustering. Since we conceived clustering as the probability that two random neighbours of a node are connected, we decided to give the same probability to all nodes. This choice corresponds to a constant clustering spectrum. This case is not the same as fixing any of the existing two clustering coefficients. In our case we enforce nodes of any degree to have the same local clustering.

In order to show the difference between fixing any of the existing clustering measures, in figure 2.2 we compare the resulting clustering spectrum and

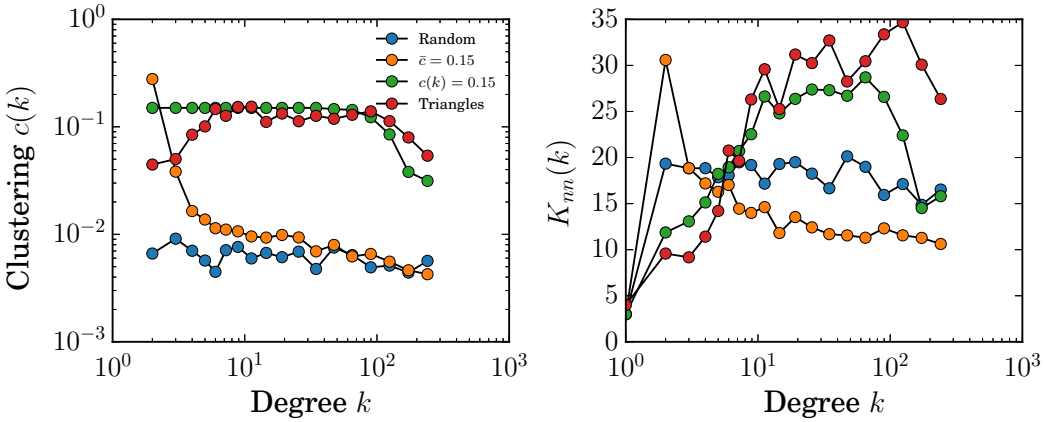


Figure 2.2: Comparison of the clustering spectrum $c(k)$ (left) and the average neighbour degree $K_{nn}(k)$ (right) as a function of the degree of four Networks. Blue: A configuration model network with 10000 nodes with a power law degree distribution with an exponent $\gamma = 2.7$, Orange: A network in which we took the random network represented in blue and fixed the clustering coefficient to $\bar{c} = 0.15$ using the RandNetGen. Green: A network in which we took the random network represented in blue and fixed a flat clustering spectrum $\bar{c}(k) = 0.15$ using the RandNetGen. Red: a Network in which we took the random network represented in blue and fixed the number of triangles to equal to the clustered network represented in green, which leads to a global clustering coefficient $\bar{C} = 0.103$.

the average neighbour degree of three networks each one with a different target clustering measure. In the three cases we start from a configuration model network with a power law degree distribution with an exponent $\gamma = 2.7$, and we make use of RandNetGen to fix different clustering measures, while preserving only the degree distribution. In the first case we fix the local clustering coefficient to $\bar{c} = 0.15$; in the second case we fix a flat clustering spectrum $c(k) = 0.15$, which gives the same local clustering coefficient as the previous case; in the third network we fix the number of triangles to be the same to the second case, which leads to the global clustering coefficient $\bar{C} = 0.103$. As we can observe in figure 2.2 the three networks have a completely different clustering spectrum and the effect that each type of clustering have on the degree correlations are very different. This fact recalls the importance of considering the hole clustering spectrum of a network instead of an average coefficient.

Another interesting application of our program is to randomize a network preserving different topological properties and observe which long-range features of the real network are well reproduced by the randomized version. This

work was done in [140] for several real networks from different natures and we concluded that, in most cases, the degree distribution, degree-degree correlation, and clustering spectrum are enough to reproduce most of the mesoscopic and macroscopic network properties. These results imply that these non-local properties do not have any specific formation mechanism, so they are just a product of the randomness, particular degree distributions and, optionally, degree correlations and clustering. This fact is in agreement with the perception that real complex networks are a product of a self-organized process in which edges are just a result of local interaction between nodes. Besides, this findings enhance the approach of this thesis, in which the degree distribution, degree correlations and clustering are the most fundamental network properties.

CHAPTER 3

Clustering of random scale-free networks

3.1 The Configuration model

Null models are critical to gauge the effect that randomness may have on the properties of systems in the presence of noise. It is therefore important to have the maximum understanding of the null model at hand, something not always easy to achieve. This is the case of the most used null model of random graphs: The configuration model (CM) [17, 30, 126, 127]. Given a real network, the configuration model preserves the degree distribution of the real network, $P(k)$, whereas connections among nodes are realized in the most random way, always preserving the degree sequence, either the real one or drawn from the distribution $P(k)$. In principle, the CM generates graphs without any type of correlations among nodes. For this reason, it is widely used in network theory to determine whether the observed topological properties of the real network might be considered as the product of some non trivial principle shaping the evolution of the system.

This program is severely hindered when the network contains nodes with degrees above the structural cut-off $k_s = \sqrt{\langle k \rangle N}$ [26], where $\langle k \rangle$ is the average degree and N the size of the network. This is the case of scale-free networks with $P(k) \sim k^{-\gamma}$, $\gamma < 3$, and a natural cut-off $k_c \sim N^{1/(\gamma-1)}$ most often found in real complex networks [135]. This apparently simple null model develops all sort of anomalous behaviours in this case, e. g., the appearance of strong non-trivial degree correlations among nodes [26, 36, 39, 142], difficulties in the sampling of the configuration space [105], or the presence of phase transitions between graphical and non-graphical phases [61], to name just a few.

Clustering –or the presence of triangles in the network– is yet another example of anomalous behaviour associated to the CM. The importance of clustering as a topological property is related to the fact that nearly all known real complex networks have a very large number of triangles whereas the CM has a vanishingly small number in the thermodynamic limit. Of course, the absence of triangles is convenient from a theoretical point of view as it allows us to use generating functions techniques to solve many interesting problems [135]. However, given

the empirical observations, it seems to be a quite unrealistic assumption. This has led to the common understanding that clustering observed in real networks cannot be explained by the CM and, thus, is the product of some underlying principle. While we fully agree with this statement, in this chapter we show that it must be taken with care. Indeed, depending on the heterogeneity of $P(k)$, the CM can generate, on average, nearly size-independent levels of clustering. Besides, in such cases, sample-to-sample fluctuations do not vanish when $N \to \infty$, meaning that the same degree sequence may generate either very high or very low levels of clustering, independently of the network size.

As mentioned in section 1.2.4, clustering can be quantified using different metrics [155]. Here, we use the average clustering coefficient $\bar{c}$, defined as the average (over nodes of degree $k \geq 2$) of the local clustering coefficient of single nodes $c_i = 2t_i/k_i(k_i-1)$, where t_i the number of triangles attached to node i. In the absence of high degree nodes, the clustering coefficient of a random graph generated by the CM is given by

$$\bar{c} = \frac{\langle k(k-1)\rangle^2}{N\langle k\rangle^3}, \tag{3.1}$$

and, therefore, vanishes very fast in the large system size [21, 135]. This is the reason why the tree-like character of networks generated by the CM has always been taken for granted. However, Eq. (3.1) is clearly incorrect when the degree distribution is scale-free with a natural cut-off $k_c \sim N^{1/(\gamma-1)}$ as it predicts a behaviour $\bar{c} \sim N^{(7-3\gamma)/(\gamma-1)}$ that diverges for $\gamma < 7/3$. Equation (3.1) fails in this case because its derivation does not account for the structural correlations among degrees of connected nodes coming from the refusal to multiple and self-connections. However, the same formula gives the correct scaling if, instead, a structural cut-off, $k_s \sim N^{1/2}$, is imposed on the degree sequence. In this case, Eq. (3.1) predicts the correct scaling $\bar{c} \sim N^{2-\gamma}$ [39]. It is then clear that the finite size scaling of the clustering coefficient in random scale-free graphs must depend on both the size of the network N and on the particular scaling of the cut-off k_c as a function of N. Here, we derive the correct scaling behaviour of the clustering coefficient for scale-free random graphs with $2 < \gamma < 3$ and any cut-off value k_c.

3.2 Maximally random graphs with expected degree sequence

The CM, as originally defined, defines a micro-canonical ensemble, in the sense that the degree of every single node is given a priori and, once the degree sequence is fully known, the network is assembled in the most random way while

preserving the degree sequence (We refer the reader to reference [60] for a method to generate such graphs without any sampling bias). However, in the case of scale-free networks, this approach resists any analytic treatment. Instead, here we adopt a different strategy and work with the canonical ensemble of the CM introduced in the previous chapter in section 2.3. In this ensemble, each node is given not its actual degree but its expected degree. This relaxes the topological conditions to close the network and opens the door to an analytic treatment.

Specifically, the model that generate a networks from the ensemble that we just defined is as follows

1. Each node is assigned a hidden variable κ drawn from the probability density $\rho(\kappa) = \propto \kappa^{-\gamma}$ with $1 \leq \kappa \leq \kappa_c$. The cut-off value κ_c is, in principle, arbitrary. However, often κ_c is the so-called natural cut-off, defined as the expected maximum value out of a sample of N random variables given from the probability density $\rho(\kappa)$. In the case of interest of a scale-free distribution, the natural cut-off scales as $\kappa_c \sim N^{1/(\gamma-1)}$.

2. Each pair of nodes is visited once and connected with probability given by

$$r\left(\frac{\kappa\kappa'}{\kappa_s^2}\right) = \frac{\kappa\kappa'}{\kappa_s^2}\left(1+\frac{\kappa\kappa'}{\kappa_s^2}\right)^{-1}. \tag{3.2}$$

   Parameter κ_s is the structural cut-off defining the onset of structural correlations, that is, nodes with expected degrees below κ_s are connected with probability $r(\frac{\kappa\kappa'}{\kappa_s^2}) \approx \frac{\kappa\kappa'}{\kappa_s^2}$ and, therefore, are uncorrelated at the level of degrees. As a consequence, the global level of correlations present in the system is controlled by the cut-off κ_c. Whenever $\kappa_c < \kappa_s$ the resulting network is fully uncorrelated whereas for $\kappa_c \geq \kappa_s$ correlations are necessary to close it. In this case, κ_s takes the form $\kappa_s = \sqrt{\frac{(\gamma-1)N(1-\kappa_c^{2-\gamma})}{(\gamma-2)\bar{k}_{min}}}$, where $\bar{k}_{min}$ is the expected minimum degree of the network. In this paper, we are interested in the range $\kappa_s \leq \kappa_c \leq N^{1/(\gamma-1)}$.

As we have seen, the average degree of a node with hidden variable κ is $\bar{k}(\kappa) \propto \kappa$. Thus, we can think of κ and $\rho(\kappa)$ as the degree and degree distribution, respectively [1].

[1] The exact form of the degree distribution was first given in [156], showing the asymptotic behaviour $P(k) \sim k^{-\gamma}$ as expected.

3.3 Clustering in maximally random graphs with expected degree sequence

Using the formalism developed in [22] (see also [20]), the local clustering coefficient of a node with hidden variable κ can be written as

$$c(\kappa) = \frac{\int_1^{\kappa_c}\int_1^{\kappa_c}\rho(\kappa')\rho(\kappa'')r\left(\frac{\kappa\kappa'}{\kappa_s^2}\right)r\left(\frac{\kappa'\kappa''}{\kappa_s^2}\right)r\left(\frac{\kappa\kappa''}{\kappa_s^2}\right)\mathrm{d}\kappa'\mathrm{d}\kappa''}{\left[\int_1^{\kappa_c}\rho(k')r\left(\frac{\kappa\kappa'}{\kappa_s^2}\right)\mathrm{d}\kappa'\right]^2}. \tag{3.3}$$

If we use the change of variables $x=\kappa'/\kappa_s$ and $y=\kappa''/\kappa_s$ we obtain

$$c(\kappa) = \frac{\int_{\frac{1}{\kappa_s}}^{\frac{\kappa_c}{\kappa_s}}\int_{\frac{1}{\kappa_s}}^{\frac{\kappa_c}{\kappa_s}}\frac{1}{(xy)^\gamma}r\left(\frac{\kappa x}{\kappa_s}\right)r\left(xy\right)r\left(\frac{\kappa y}{\kappa_s}\right)\mathrm{d}x\mathrm{d}y}{\left[\int_{\frac{1}{\kappa_s}}^{\frac{\kappa_c}{\kappa_s}}x^{-\gamma}r\left(\frac{\kappa x}{\kappa_s}\right)\mathrm{d}x\right]^2}. \tag{3.4}$$

The average clustering coefficient is computed from $c(\kappa)$ as $\bar{c}=\int_1^{\kappa_c}\rho(\kappa)c(\kappa)\mathrm{d}\kappa$ [2]. However, $c(\kappa)$ is a bounded monotonously decreasing function and so its major contribution to $\bar{c}$ comes from nodes with small degree, i. e., low κ [39]. Therefore, to find the correct scaling behaviour it suffices to evaluate $c(\kappa)$ in the domain $\kappa << \kappa_s$. In this case, the maximum value within the domain of integration $[1/\kappa_s,\kappa_c/\kappa_s]$ of the arguments $\kappa x/\kappa_s$ and $\kappa y/\kappa_s$ in Eq. (3.4) is of order $\mathcal{O}(\kappa_c/\kappa_s^2)$, which goes to zero in the thermodynamic limit. We can, thus, approximate $c(\kappa)$ as

$$c(\kappa) \approx \frac{(\gamma-2)^2}{\kappa_s^{2(\gamma-2)}(1-\kappa_c^{2-\gamma})^2}\int_{\frac{1}{\kappa_s}}^{\frac{\kappa_c}{\kappa_s}}\int_{\frac{1}{\kappa_s}}^{\frac{\kappa_c}{\kappa_s}}\frac{(xy)^{2-\gamma}}{1+xy}\mathrm{d}x\mathrm{d}y, \tag{3.5}$$

which becomes independent of κ.

To solve this integral we use the transcendent Lerch function $\Phi(z,a,b)$ [89]. We use the following identity

$$\int_0^a dx\int_0^b dy\frac{(xy)^{2-\gamma}}{1+xy} = (ab)^{3-\gamma}\Phi(-ab,2,3-\gamma) \tag{3.6}$$

for $2<\gamma<3$ and $a,b>0$, which allows us to write Eq. (3.5) as

$$c(\kappa) \approx \frac{(\gamma-2)^2}{\kappa_s^{2(\gamma-2)}(1-\kappa_c^{2-\gamma})^2}\left[\left(\frac{\kappa_c}{\kappa_s}\right)^{2(3-\gamma)}\Phi\left(-\left(\frac{\kappa_c}{\kappa_s}\right)^2,2,3-\gamma\right)\right. \tag{3.7}$$

[2] Notice that since κ is only the expected degree, a node with, for instance, expected degree $\kappa=1$ may end up with an actual degree above 1 and vice versa. This implies that all values of κ contribute to the global clustering of the network and, thus, the domain of integration is $[1,\kappa_c]$

$$-2\left(\frac{\kappa_c}{\kappa_s^2}\right)^{3-\gamma}\Phi\left(-\frac{\kappa_c}{\kappa_s^2},2,3-\gamma\right)+\frac{1}{\kappa_s^{6-2\gamma}}\Phi\left(-\frac{1}{\kappa_s^2},2,3-\gamma\right)\Bigg].$$

The first argument of the second and third transcendent Lerch functions in this equation goes to zero in the thermodynamic limit because $\kappa_c << \kappa_s^2 \sim N$. However, the argument of the first Lerch function diverges unless $\kappa_c \sim \kappa_s$. Unfortunately, there is not known asymptotic behaviour for the Lerch function for diverging arguments. To overcome this problem, we use the integral representation of function $\Phi(-z^2,2,3-\gamma)$,

$$\Phi(-z^2,2,3-\gamma)=\int_0^\infty \frac{xe^{-(3-\gamma)x}}{1+z^2e^{-x}}dx. \tag{3.8}$$

The domain of integration in Eq. (3.8) can be separated in the sub-domains $[2,2\ln z]$ and $(2\ln z,\infty)$ such that function $(1+z^2e^{-x})^{-1}$ can be expanded as a converging Taylor series in each sub-interval. Once this trick is used, it is easy to derive the following identity

$$\Phi(-z^2,2,3-\gamma)=z^{-2(3-\gamma)}\left[2\psi(\gamma)\ln z+\theta(\gamma)\right]+\frac{1}{z^2}\Phi\left(-\frac{1}{z^2},2,\gamma-2\right). \tag{3.9}$$

Notice that this expression has a well defined behaviour when $z \gg 1$. Plugging this expression into Eq. (3.7) we obtain

$$c(\kappa)\approx\frac{(\gamma-2)^2}{\kappa_s^{2(\gamma-2)}(1-\kappa_c^{2-\gamma})^2}\left[2\psi(\gamma)\ln\left(\frac{\kappa_c}{\kappa_s}\right)+\theta(\gamma)+\left(\frac{\kappa_s}{\kappa_c}\right)^{2(\gamma-2)}\Phi\left(-\left(\frac{\kappa_s}{\kappa_c}\right)^2,2,\gamma-2\right)\right. \tag{3.10}$$
$$\left.-2\left(\frac{\kappa_c}{\kappa_s^2}\right)^{3-\gamma}\Phi\left(-\frac{\kappa_c}{\kappa_s^2},2,3-\gamma\right)+\frac{1}{\kappa_s^{6-2\gamma}}\Phi\left(-\frac{1}{\kappa_s^2},2,3-\gamma\right)\right]$$

where

$$\psi(\gamma)=\Phi(-1,1,3-\gamma)+\Phi(-1,1,\gamma-2),$$
$$\theta(\gamma)=-\pi^2\cot\pi\gamma\csc\pi\gamma.$$

This expression, although involved at first glance, it is convenient because in the range $\kappa_s \le \kappa_c \ll \kappa_s^2$ the arguments of the three transcendent Lerch functions in it go to 0^- in the limit $\kappa_s \to \infty$, in which case we know that $\Phi(-z^2,a,b) \sim b^{-a}$ for $z \to 0$. We then find the asymptotic behaviour

$$c(\kappa)\sim\frac{(\gamma-2)^2}{\kappa_s^{2(\gamma-2)}}\begin{cases}\theta(\gamma)+\Phi(-1,2,\gamma-2) & \kappa_c=\kappa_s\gg 1\\ 2\psi(\gamma)\ln\left(\frac{\kappa_c}{\kappa_s}\right) & \kappa_c\gg\kappa_s\gg 1.\end{cases} \tag{3.11}$$

The first line in this equation recovers the result found in [39] for scale-free networks without structural correlations –$c(\kappa) \sim N^{2-\gamma}$ when $\kappa_s \sim N^{1/2}$– whereas the

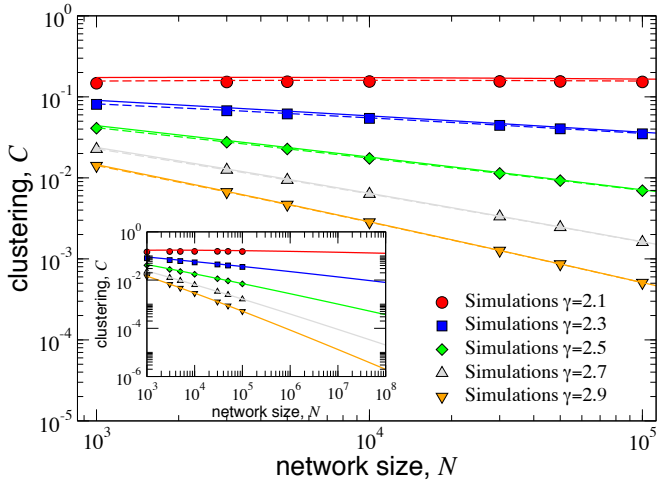


Figure 3.1: Clustering coefficient as measured in numerical simulations for different values of γ and size N with $\bar{k}_{min} = 2$ and $\kappa_c = N^{1/(\gamma-1)}$. Each point is an average over 10^4 different network realizations. Dashed lines are the numerical solution of Eq. (3.4) and solid lines are the approximate solution given by Eq. (3.10). The inset shows an extrapolation up to size $N = 10^8$ using Eq. (3.10).

second line predicts $c(\kappa) \sim N^{2-\gamma} \ln N$ when $\kappa_c \sim N^{1/(\gamma-1)}$, which corrects the incorrect scaling behaviour predicted by Eq. (3.1) in this case. Interestingly, this scaling is different from the one found for a model of growing random scale-free graphs [12], again making evident the difference between equilibrium and non-equilibrium models of random graphs [63].

Figure 3.1 shows a comparison between numerical simulations, the numerical solution of Eq. (3.4), and the approximate solution given by Eq.(3.10), showing a very nice agreement. Interestingly, for $\gamma = 2.1$, clustering remains nearly constant in the range of sizes $10^3 - 10^5$ and even increases slightly for small sizes. This is a consequence of the slow decay of the term $\kappa_s^{2(2-\gamma)}$ combined with the diverging logarithmic term in the numerator and functions $\psi(\gamma)$ and $\theta(\gamma)$, which diverges in the limit $\gamma \to 2$. In the inset of Fig. 3.1, we show the extrapolation of the clustering coefficient for sizes up to 10^8 evaluated with Eq. (3.10). In the case of $\gamma = 2.1$, this figure makes evident the extremely slow decay –nearly absent– with the system size. This implies that, in practice, clustering cannot be removed from the network even in very large networks when $\gamma \approx 2$. It is, thus, not clear whether the tree-like approximation, customarily used to solve problems on random graphs, can be applied in this case. In this situation, one should use alternative approaches, like the one developed in [159]. These results are particularly relevant due to the abundance of real networks with values of $\gamma \approx 2$. It is also interesting to study the behaviour of clustering as a function of γ for

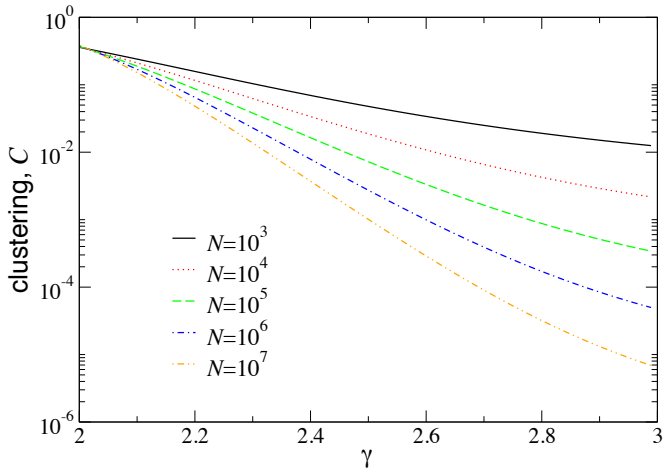


Figure 3.2: Clustering coefficient as a function of γ for different network sizes. Curves are evaluated from Eq. (3.10) with $\bar{k}_{min} = 2$ and $\kappa_c = N^{1/(\gamma-1)}$.

a fixed network size. Figure 3.2 shows this behaviour for different values of N, confirming the results found in Fig. 3.1. Clustering increases as γ decreases and converges to a constant and size independent value at $\gamma = 2$.

Up to this point, we have been concerned only with the ensemble average of the clustering coefficient. However, the CM ensemble shows strong sample-to-sample fluctuations. Figure 3.3 shows the probability density function of the clustering coefficient obtained out of a sample of 10^4 different networks generated by the canonical version of the CM. As it can be observed, clustering may take values in the range $[0.05, 0.25]$ quite easily. Figure 3.3 also shows the standard deviation σ_C as a function of network size and for different values of γ. In all cases, fluctuations decay as a power law of the system size, $\sigma_C \sim N^{-z}$, with an exponent $z < 1$. Interestingly, for $\gamma = 2.1$, the exponent z takes a very small value ($z \approx 0.1$) that, when combined with the behaviour of C as a function of N results in a coefficient of variation nearly constant. This implies that, in this range of values of γ, clustering is *de facto* a size-independent but non self-averaging property. That is, a single network instance is not a good representative of the ensemble even for very large network sizes.

3.4 Discussion

The presence of triangles in real networks play an important role in many processes taking place on top of them, e. g. , percolation phenomena, epidemic spreading, synchronization, etc. It is, therefore, important to have full control over the most simple network ensembles that are used as null models to assess

the presence of underlying principles shaping the topology of the system.

Here we have found the correct scaling behaviour of the clustering coefficient of the ensemble of scale-free random graphs with $2 < \gamma < 3$. Interestingly, for values of the exponent $\gamma \approx 2$, clustering remains nearly constant up to extremely large network sizes. However, in this case, clustering is not self-averaging. This means that when comparing real networks against the CM, it is not enough to generate a single instance network, as it may result in either a very low or high level of clustering even for very large network sizes. These results are particularly important as the exponent $\gamma \approx 2$ seems to be –for yet unknown reasons– the rule rather than the exception in real systems.

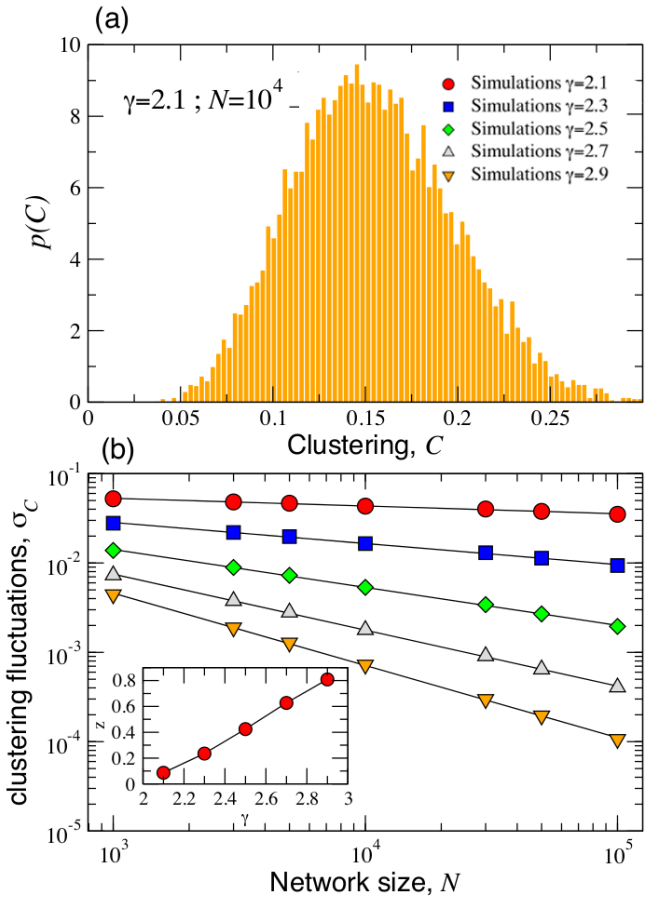


Figure 3.3: Sample to sample fluctuations. Plot (a) shows the probability density function of the clustering coefficient as obtained from 10^4 network realizations for $\bar{k}_{min} = 2$, $\kappa_c = N^{1/(\gamma-1)}$, $\gamma = 2.1$, and $N = 10^4$. Plot (b) shows the standard deviation of this pdf for different values of γ as a function of the network size. Solid lines are power law fits of the form $\sigma_C \sim N^{-z}$. The exponent z is shown in the inset.

CHAPTER 4

Global organization of clustering in complex networks

4.1 How are the triangles organized?

The architecture of real complex systems lies between order and disorder, although its precise location is quite difficult to determine. On the one hand, disorder in complex networks is manifested by the small-world effect [170] and a highly heterogeneous degree distribution [9], both properties commonly present in real complex networks [63, 135]. On the other hand, order is manifested by the presence of triangles –or clustering– representing three point correlations in the system. Indeed, the very concept of order is typically related to the existence of a metric structure in the system which, from the network perspective, is captured by clustering, the smallest network motif able to encode the triangle inequality. Yet, unlike the small-world effect and the heterogeneity of nodes' degrees, clustering is not an emergent property spontaneously generated by paradigmatic connectivity principles such as preferential attachment. Besides, the most popular network model, the configuration model, we have seen that, although it can show high levels of clustering in some situations, its not enough to reproduce the ones observed in real networks. Therefore, clustering calls for specific mechanisms for explaining its emergence, thus giving important insights into the nature of network formation and network evolution.

However, the effects of clustering on the structural and dynamical properties of networks have not yet been conclusively elucidated. In fact, several studies have reported apparently contradictory results concerning the effects of clustering on the percolation properties of networks and little is known on its effects on dynamical processes running on networks [84, 85, 86, 134, 153, 154, 167]. This is further hindered by the technical difficulties of any analytical treatment. Indeed, the presence of strong clustering invalidates, in general, the "locally tree-like" assumption used in random graphs, leaving little room for any theoretical study.

In an effort to overcome these problems, a new class of clustered network models has been proposed [83, 84, 85, 86, 98, 122, 134, 167]. These are based on the idea of introducing clustering in the network by means of cliques of dif-

ferent sizes. While different models have different rules to match these cliques to close the network, they are all based on the same principles used in the classical configuration model to generate random graphs with a given degree sequence. In this way, the resulting clustered graph is embedded in another graph that is locally tree-like, thus allowing for an analytical treatment. With these approaches, it is possible to generate networks with a given degree distribution $P(k)$ and degree-dependent clustering coefficient $\bar{c}(k)$, or clustering spectrum.

While this is indeed a fair approach to the problem, triangles generated by these models are arranged in a very specific way, with strong correlations between the properties of adjacent edges. In some sense, we can consider this class of models as generators of maximally ordered clustered graphs. At the other side of the spectrum, we can define an ensemble of maximally random clustered graphs such that correlations among adjacent edges are the minimum needed to conform with the degree-dependent clustering coefficient, but no more. These two types of models define –in a non-rigorous way– two extremes of the phase space of possible graphs with given $P(k)$ and $\bar{c}(k)$. A simple question arises then: where are real networks positioned in this phase space?

To give an answer to this question, we need to go beyond the local properties of networks and to study their global organization. In this chapter, we study the global structure of clustering in real networks and compare them with the global structure of clustering induced by the two types of models with identical local properties. More specifically, we analyse the organization of real and model networks into m-cores, defined as maximal sub-graphs with edges participating at least in m triangles, that is able to distinguish between hierarchical and modular architectures. Interestingly enough, real networks tend to be closer to maximally random clustered graphs, although clear differences are evident.

In this chapter, we analyse three real paradigmatic networks from different domains: the Internet at the Autonomous System level [28], the web of trust of the Pretty Good Privacy protocol (PGP) [25], and the metabolic network of the bacterium *E. coli* [158]. However, the results obtained here also hold for a wide spectrum of systems [50] (See the appendix A.1 for an extensive description of these real networks).

4.2 Clustered network models

One of the best clique-based models to generate maximally ordered clustered networks is the one introduced by Gleeson in [84]. In this model, nodes belong to single cliques and are also given a number of connections outside their cliques. Then, cliques are considered as super-nodes, each with an effective degree given by the sum of all the external links of the members of the clique, and

connected using the standard configuration model. The input of the model is the joint distribution $\gamma(c,k)$, defined as the probability that a randomly chosen node has degree k and belongs to a clique of size c. Both the degree distribution and the degree-dependent clustering coefficient are related to function $\gamma(c,k)$. Therefore, by properly choosing its form, it is possible to match the desired degree distribution and clustering. Note, however, that since we start with cliques and not nodes, the number of nodes and their actual degrees are not fixed *a priori*. As a consequence, in finite heterogeneous networks, there may be some unavoidable discrepancies between real and random versions of the network. Hereinafter, we denote this model as "clique-based model" (CB).

On the other, we generate maximally random clustered networks as an ensemble of exponential graphs as introduced in section 2.5 with Hamiltonian

$$H = \beta \sum_{k=k_{min}}^{k_c} |\bar{c}^*(k) - \bar{c}(k)|, \tag{4.1}$$

where k_{min} and k_c are the minimum and maximum degrees of the network, $\bar{c^*}(k)$ is the target degree-dependent clustering coefficient, $\bar{c}(k)$ is the one corresponding to the current state of the network and β is the Lagrange multiplier. Therefore, starting from a given real network and after an initial randomization, this Hamiltonian is minimized by means of simulated annealing coupled to a Metropolis rewiring scheme until the current clustering is close enough to the target one. Here we use both rewiring schemes. The one that only preserves the degrees of nodes and also the one that preserve both the degree distribution and the joint degree-degree distribution of connected nodes, $P(k,k')$, so that degree-degree correlations are fully preserved. As we have shown in section 2.2, this network model generates network with a given constrain, here the $\bar{c}(k)$, and is maximally random to all other respects. Hereinafter, we denote these models as "maximally random models" (MR). We would like to stress that, even though there are many models of exponential random graphs generating clustered graphs [73, 76, 123], none of them reproduces the actual clustering spectrum as a function of node degree. In this sense, our maximally random model gets closer to real networks.

Notice that none of the random models used in this chapter enforces global connectivity of the network in a single connected component. Therefore, the number of disconnected components and the size of the giant (or largest) component must be considered as predictions of the models, which can be readily compared to those of real networks. In Table 4.1, we show this comparison with the networks analysed here. Quite remarkably, in the case of the Internet, MR models predict the existence of, basically, a single connected component, as it is also observed in the real network. On the other hand, the CB model generates

Table 4.1: Statistics of real networks and their random counterparts. N is the number of nodes, E is the number of edges, $\bar{c}$ is the average clustering coefficient averaged only over nodes with degrees $k \geq 2$. We also show the number of disconnected components (clusters) and the relative size of the giant component. Error bars are computed as the standard deviation of the corresponding metric as obtained from a sample of 10 network realizations. Figures without errors did not show any significant difference between different samples.

	N	E	$\bar{c}$	# of clusters	Giant component
Internet	23752	58416	0.61	3	99.98%
Internet clique-based model	23800±200	50000±10000	0.62±0.01	2200±400	(75±4)%
Internet random $\bar{c}(k)$	23752	58416	0.61	16±4	(99.84±0.06)%
Internet random $\bar{c}(k)$, $P(k,k')$	23752	58416	0.61	4±1	(99.96±0.02)%
PGP	57243	61837	0.50	16221	18.65%
PGP clique-based model	62000±1000	57200±200	0.506±0.005	13700±200	(37±1) %
PGP random $\bar{c}(k)$	57243	61837	0.487±0.001	15550±60	(21.3±0.4)%
PGP random $\bar{c}(k)$, $P(k,k')$	57243	61837	0.493±0.001	15810±20	(22.3±0.3)%
E. Coli	1010	3286	0.48	2	99.8%
E. Coli clique-based model	1010±40	3300±700	0.51±0.01	7±3	(97.9±0.6) %
E. Coli random $\bar{c}(k)$	1010	3286	0.48	2.2±0.9	(99.7±0.3)%
E. Coli random $\bar{c}(k)$, $P(k,k')$	1010	3286	0.48	7±2	(98.2±0.6)%

a very large number of disconnected components and a giant component significantly smaller than the real one. Even more surprising are the results for the PGP web of trust. The real network is fragmented into a large number of small components whereas its giant component occupies around 18% of the network. All models generate a similar number of disconnected components. However, the relative size of the giant component is very well reproduced by MR models, whereas the CB model predicts a giant component twice as large. In the case of the metabolic network of the bacterium *E. coli*, all models predicts the existence of a single connected component, in good agreement with the real network.

4.3 Revealing network hierarchies: k-cores and m-cores

Real heterogeneous networks are typically hierarchically organized. One of the most useful tools to uncover such hierarchies is the k-core decomposition [65]. Given a network, its k-core is defined as the maximal subgraph such that all nodes in the subgraph have at least k connections with members of the subgraph. This defines a hierarchy of nested sub-graphs, where the 1-core contains the 2-core, which in turn contains the 3-core and so on until the maximum k-

core is reached. Nodes belonging to the k-core but not to the $(k+1)$-core are said to have coreness k. Real networks often show a deep and complex k-core structure, as made evident by tools such as LaNet-Vi [15]. However, even though clustering has been shown to induce strong k-core hierarchies [154], the k-core *per se* does not include any information about clustering and, thus, cannot discriminate well between two networks with different global organization of clustering but with the same clustering coefficient.

To overcome this problem, the concept of k-core has been remodelled to account for clustered networks. A key ingredient throughout the paper is the concept of edge multiplicity m, defined as the number of distinct triangles going through an edge [149, 155, 171]. All edges belonging to a clique of size n have identical multiplicity $n-2$ whereas an edge connecting two cliques has zero multiplicity. Therefore, strong correlations between the multiplicities of adjacent edges indicate that triangles are arranged in a clique-like fashion whereas a weaker correlation indicate a random distribution of triangles. It is therefore clear that, in order to uncover the global organization of triangles in a network, it is necessary to understand the organization of the multiplicities of their edges. This can be achieved with the m-core, defined as the maximal subgraph such that all its edges have, at least, multiplicity m within it. This concept was developed in [91, 152] under the name of k-dense decomposition. The edges in a k-dense graph have multiplicity $m = k-2$. Because of this, we prefer the notion of m-core, which is directly related to the multiplicity: an edge belongs to the m-core if its multiplicity within the m-core is, at least, m. A node belongs to the m-core if at least one of its edges belongs to it. A node belonging to the m-core but not to the $(m+1)$-core is said to have m-coreness m. As in the case of the k-core, the m-core defines a set of nested sub-graphs whose properties informs us about the global organization of triangles in the graph. Figure 4.1 shows an example of a simple network and its m-core structure.

In the case of the k-core, the internal average degree within each subgraph grows as k is increased. As a consequence, it is very unlikely that the $(k+1)$-core is fragmented in different components if the k-core is connected. Therefore, the main interest of the k-core decomposition is focused on the size of the giant k-core and the maximum coreness of the system. The situation is completely different in the case of the m-core. This is so because of a weaker correlation between m-coreness of a node and its degree [139]. In fact, the m-core decomposition is able to distinguish between a strong hierarchical structure –when m-cores do not fragment into smaller components– from a highly modular architecture –when m-cores are always fragmented. In this case, the quantities of interest are, besides the size of the giant m-core and the maximum m-coreness, the number of components as a function of m.

Figures 4.2, 4.3, and 4.4 show a comparison of the k-core and m-core de-

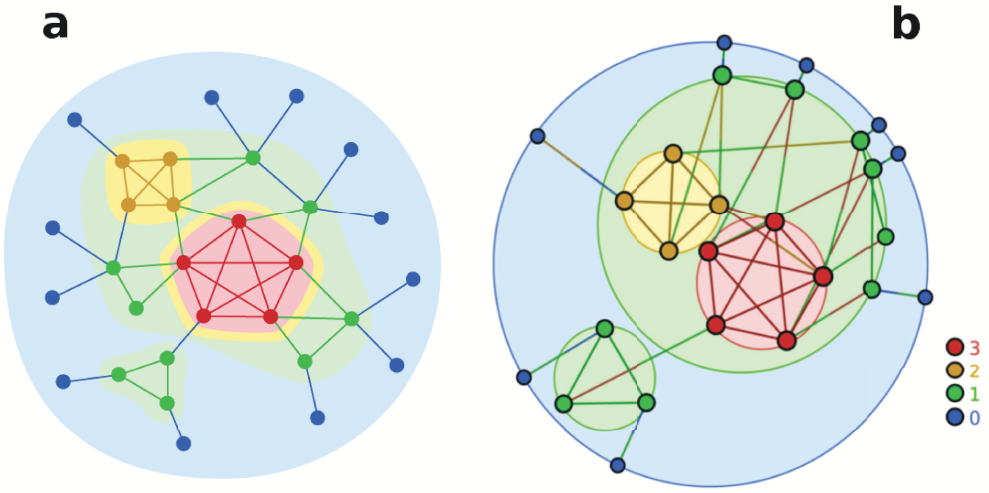


Figure 4.1: **m-cores decomposition and its visualization.** The example network in **a** is coloured according to the m-coreness of nodes and edges. Nodes and edges coloured in blue belong to the $m0$-core but not to the $m1$-core. Nodes and edges coloured in green belong to the $m1$-core but not to the $m2$-core, etc. The same structure is represented in **b** with the visualization tool described in the main text. The outermost circle in blue represents the $m0$-core, with nodes of m-coreness 0 located in its perimeter. The $m1$-core –which is contained within the $m0$-core– is fragmented in two disconnected components, which are represented as two non-overlapping circles within the outermost one and with nodes of m-coreness 1 located in their perimeters. The larger of these two components is further fragmented in two disconnected components representing the $m2$-core and $m3$-core. The angular positions of nodes in each circumference are chosen to minimize the angular separation with their neighbours in different layers. Notice that in this representation, each edge is coloured with two colors, corresponding to the colors of the m-coreness of the nodes at the end of the edge but in reverse order. In this way, it is possible to visualize easily connections between different layers. See [15] for further details of the visualization.

compositions between real networks and their random equivalents. As it can be observed in the top plots of these figures, all models do a reasonably good job at reproducing both the k-core structure and the distribution of edge multiplicities, even though MR models are clearly better than the CB one. However, there are important differences in the m-core decomposition. While both versions of MR models reproduce well the giant m-core, the maximum m-coreness, and the number of components as a function of m of all the studied networks, the CB model overestimates the size and number of components in the case of the Internet and underestimate the size of giant m-cores in the PGP web of trust. In the case of the metabolic network, MR models reproduce well its entire m-core structure. The CB model, on the other hand, does not capture well the m-core decomposition. Even though the CB network is originally connected, it fragments into a large number of disconnected components already at the $m1$-core and keeps fragmenting at each level almost up to the largest m-core, which is also three times larger than the real one.

4.4 m-core visualization

The m-core decomposition is actually much richer and complex than what Figs. 4.2, 4.3, and 4.4 show. Certainly, the m-core decomposition can be represented as a branching process that encodes the fragmentation of m-cores into disconnected components as m is increased. The tree-like structure of this process informs us about the global organization –for instance hierarchical vs. modular– of clustering in networks. To visualize this process we developed LaNet-vi 3.0, a modified version of LaNet-vi, originally designed to visualize the k-core structure of a network [15], but now extended to include the m-core decomposition. We have made our code publicly available to the scientific community on SourceForge [16]. In short, the old LaNet-vi tool evaluates the coreness of all nodes of the network and arranges them in a plane following the hierarchy induced by the k-cores, so that nodes with high coreness are placed at the center of the figure whereas nodes with lower coreness are located around nodes with higher coreness in an onion-like shape. The major modification in LaNet-vi 3.0 with respect to the visualization mode in the previous version concerns the representation of disconnected components. If the network forms a single connected component, nodes with m-coreness 0 are arranged in the outermost circle of the representation. Whenever the $m1$-core is fragmented into several components, these are arranged in separate and non-overlapping disks within the circle of m-coreness 0, with nodes of m-coreness 1 placed at the edge of their corresponding disk. The process is repeated for each disconnected component with the $m2$-core, $m3$-core, etc., until the maximum m-coreness present

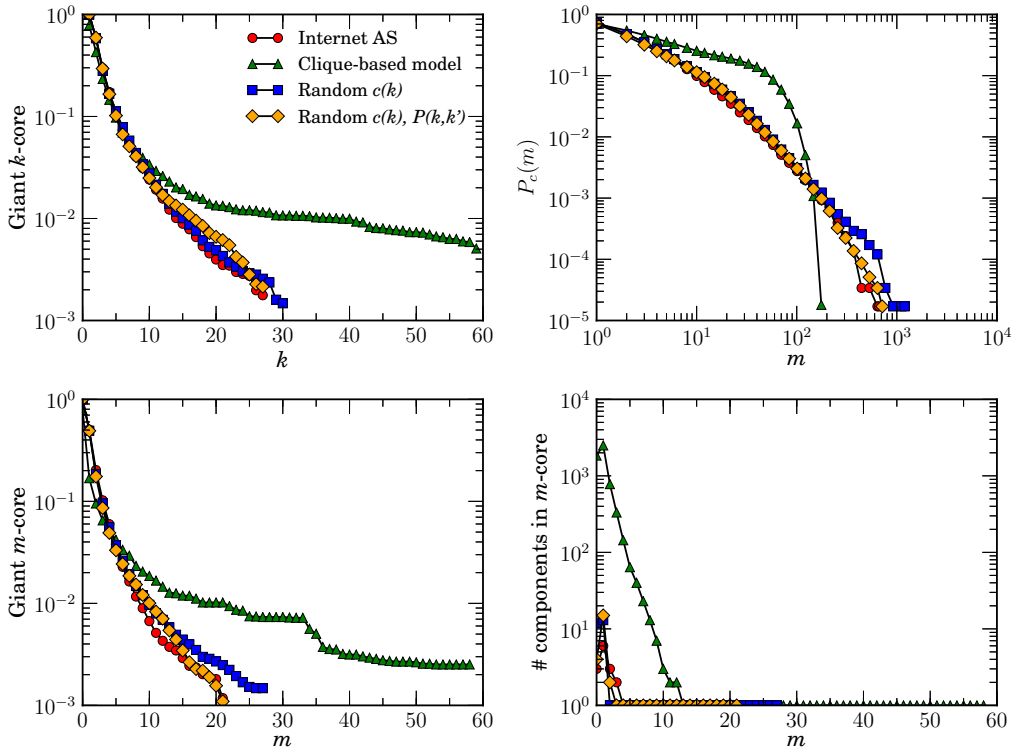


Figure 4.2: **Measuring hierarchies in real and random networks.** Comparison of the k-core and m-core decompositions between the real Internet AS network, the clique based model, and maximally random models. "Random $c(k)$" stands for the maximally random model with a fixed degree distribution and clustering spectrum $c(k)$. "Random $c(k)$, $P(k,k')$" stands for the maximally random model that preserves also the degree-degree correlation structure of the real network. The top left plot shows the relative size of the giant k-core as a function of k. Top right plot shows the complementary cumulative distribution of edge multiplicities. Bottom left plot shows the relative size of the giant m-core as a function of m. Finally, the bottom right plot shows the number of components in the m-core as a function of m.

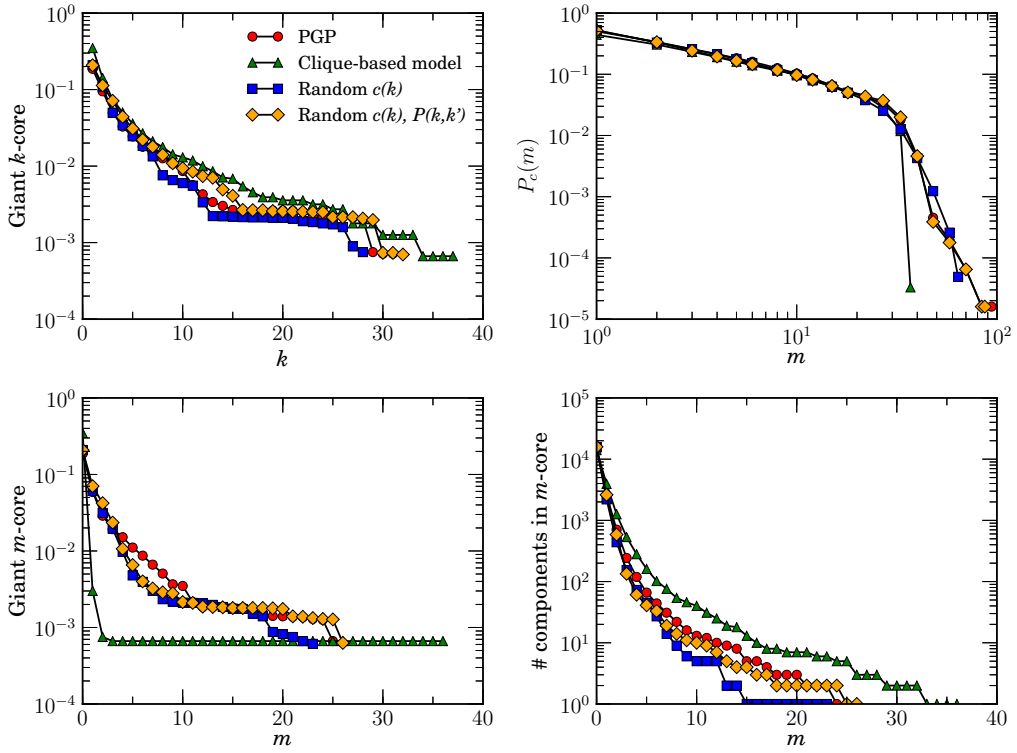


Figure 4.3: **Measuring hierarchies in real and random networks.** The same as in Fig. 4.2 but for the PGP web of trust.

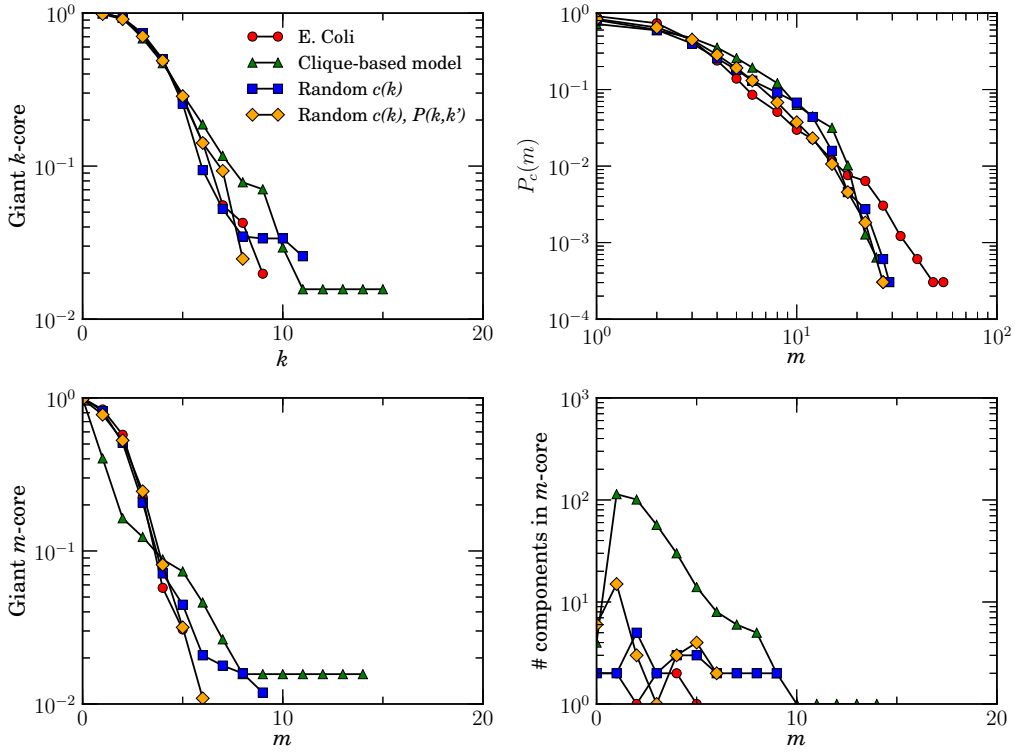


Figure 4.4: **Measuring hierarchies in real and random networks.** The same as in Fig. 4.2 but for the *E. Coli* metabolic network.

in the network is reached. The size of each disk is proportional to the logarithm of the number of nodes in the component. In this way, it is possible to visualize simultaneously all the information encoded in the m-cores so that different networks can be easily compared (see the right plot in Fig. 4.1 for a simple example). When the original network is already fragmented (like in the PGP web of trust, for instance), we first proceed to arrange disconnected components in non overlapping disks within the outermost disk, that in this case does not have any node in its perimeter.

Figures 4.5, 4.6, and 4.7 show the visualization of m-cores of real networks and their random equivalents (visualizations of MR models are shown only for $P(k)$ preserving rewiring). In the case of the Internet graph, the m-core visualization reveals a strongly hierarchical structure, where each layer is contained within the previous layer and where connections are mainly radial, with nodes with low m-coreness connected to nodes with higher m-coreness and very few connections between nodes in the same layer. Interestingly, this type of structure is also revealed in recent embeddings of the Internet graph into the hyperbolic plane [28]. This structure is very well reproduced by MR models, as it can be seen in the left bottom plot of Fig. 4.5, but not by the CB model, which generates a highly modular and non-hierarchical structure. The case of the web of trust of PGP is particularly interesting. Figure 4.6 reveals a mixture of a modular structure, with a strong fragmentation for all values of m –as one would expect for a social network–, and a hierarchical structure, revealed by the existence of a persistent giant m-core and a large number of layers. Again, this structure is very well reproduced by MR models whereas the CB model generates a very flat modular structure without any hierarchy. Finally, the metabolic network is also strongly hierarchical, although due to the small network size the number of layers is relatively small. MR models reproduce very well its structure whereas the CB model does not generate any hierarchy.

4.5 Discussion

The results presented in this chapter indicate that, in agreement with previous studies [74, 95], the degree distribution $P(k)$ and clustering spectrum $\bar{c}(k)$ are the main contributors to the global organization of the majority of real networks, which are close to maximally random once these properties are fixed. This supports the idea that most real networks are the result of a self-organized process based on local optimization rules, in contrast to global optimization principles, that yield a hierarchical organization that cannot be reproduced by maximally ordered clustered models. Besides, the strong clustering observed in real networks, supports also the idea that such local principles are related to a similarity

measure among nodes of the network that can be quantified by an underlying metric structure [27, 28, 108, 141, 156, 158]. On the other hand, global optimization principles are necessarily present, for instance, in power grids, where they induce topologies that are very different from what one would expect at random. This is made evident by its m-core decomposition (see Appendix A.1.4). In this case, even thought the m-core structure is not very deep, it is very different from any of the random models, which generate highly unstructured m-cores. Therefore, the m-core decomposition along with its visualization tool can help us to find the true mechanisms at play in the formation and evolution of real networks.

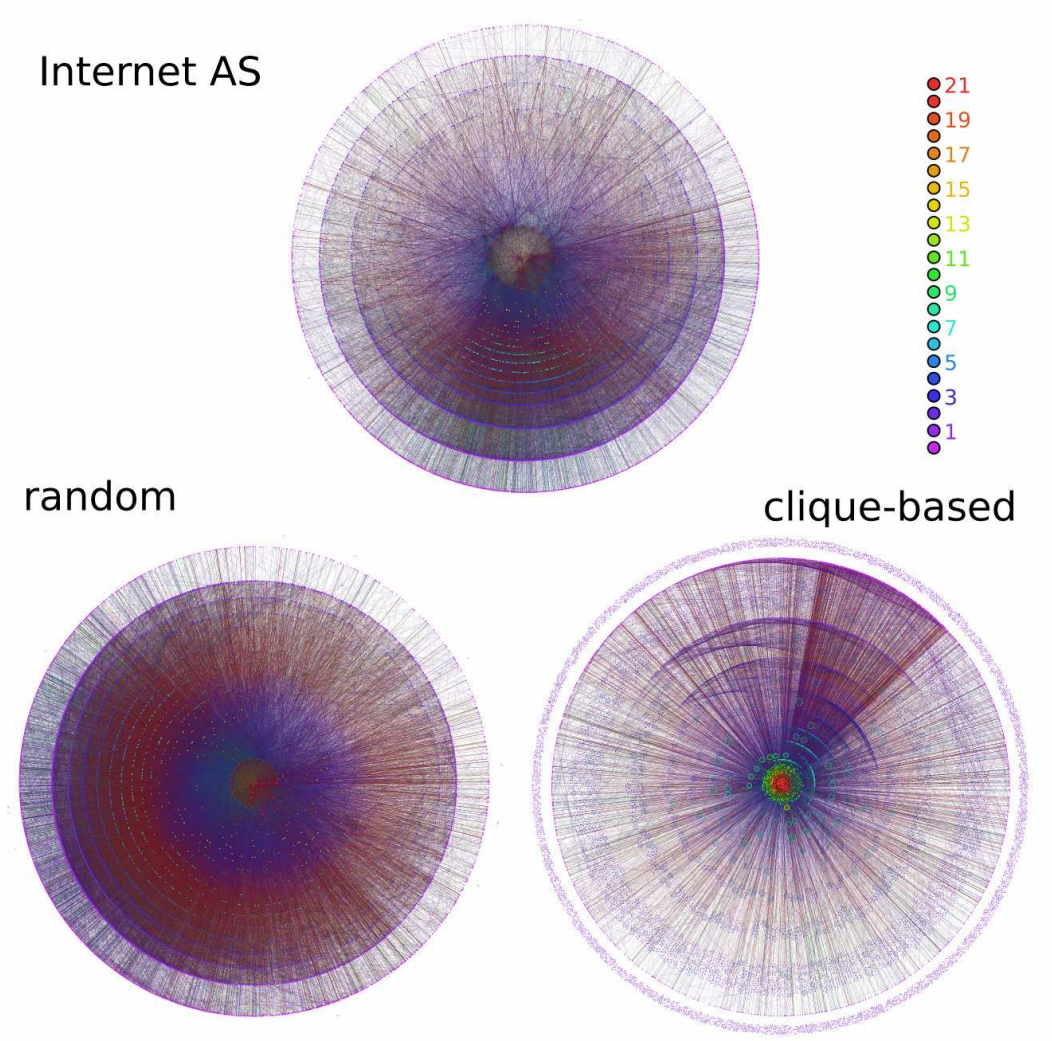


Figure 4.5: **Visualizing m-cores.** m-core decomposition of the Internet AS network and its random versions. The MR version shown on the bottom left plot of the figure corresponds to the "Random $c(k)$" model, that is, with the rewiring scheme that does not preserves degree-degree correlations. The latter case is always closer to the real network. The color code is determined by the real network and kept the same in its random versions. However, layers in random networks above the maximum m-coreness of the real network are coloured all in red. Maximum m-coreness for the MR and CB models are 27 and 58, respectively.

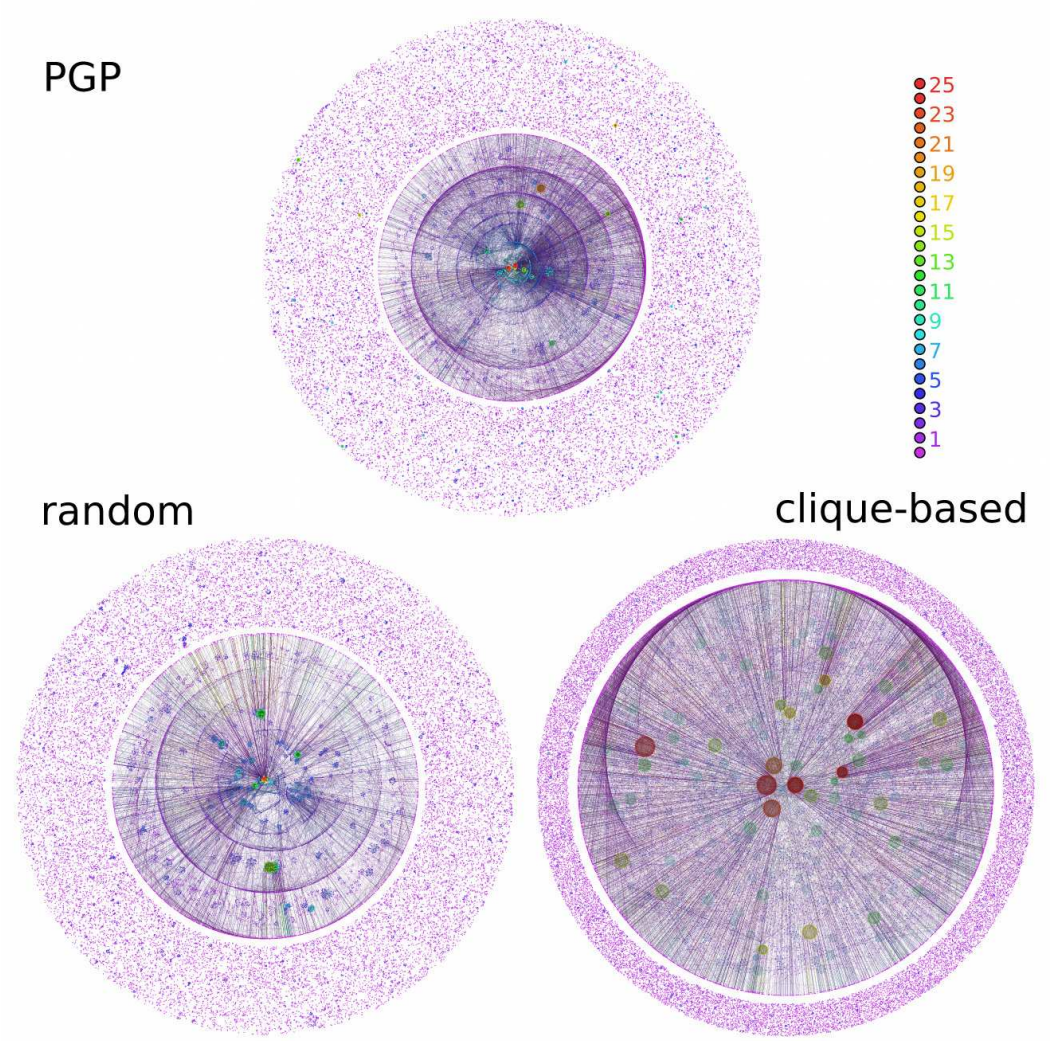


Figure 4.6: **Visualizing m-cores.** The same as in Fig. 4.5 for the PGP network and its random versions. Maximum m-coreness for the MR and CB models are 23 and 36, respectively.

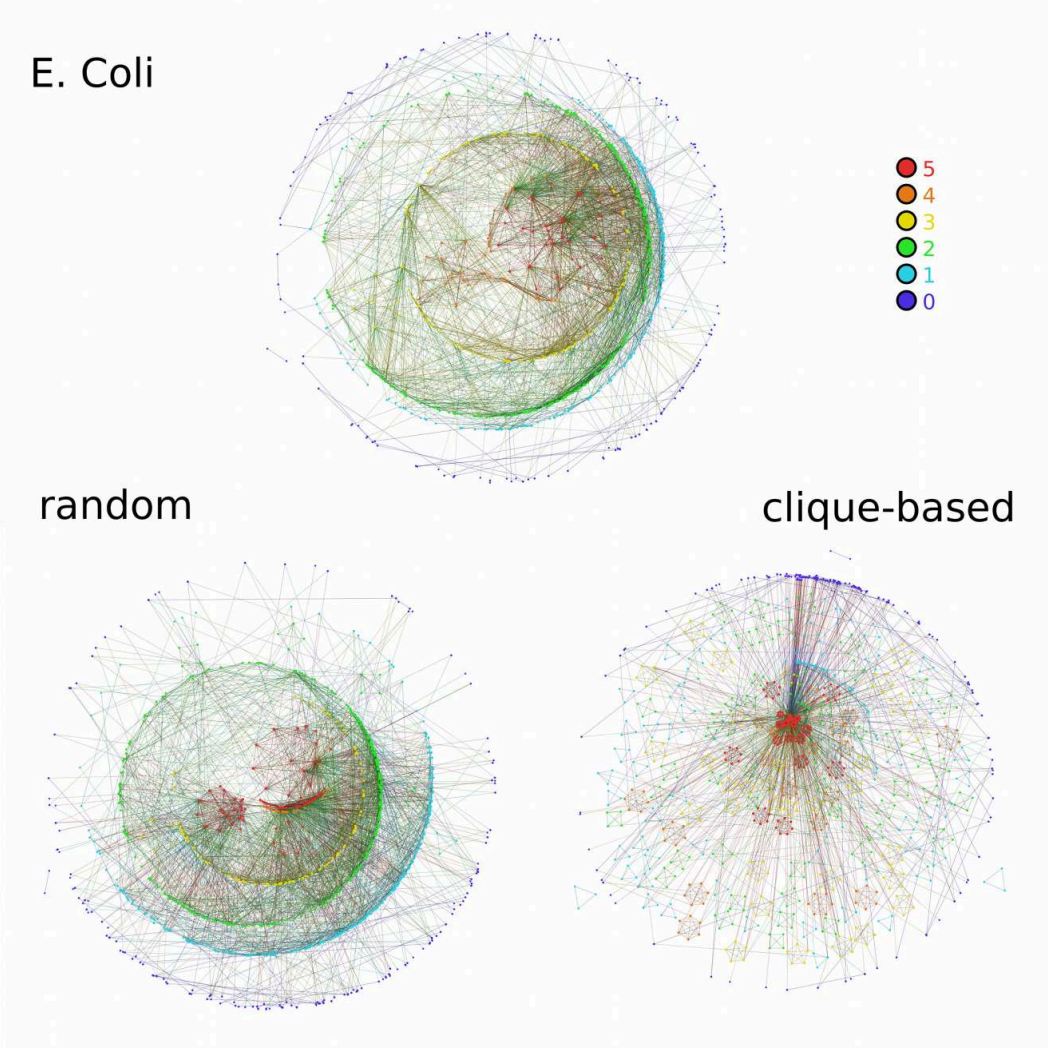


Figure 4.7: **Visualizing m-cores.** The same as in Fig. 4.5 for the *E. Coli* metabolic network and its random versions. Maximum m-coreness for the MR and CB models are 9 and 14, respectively.

CHAPTER 5

Bond percolation

So far we have looked at the relation of clustering with other topological properties and its effect on the global structure of networks. In chapter 3 we studied, both analytically and empirically, the clustering of random scale-free networks revealing the relation between the degree distribution and clustering. Then, in chapter 4, we unveiled the global organization of triangles in real networks which we showed that is close to maximally random. Therefore, in contrast to other models in which triangles are distributed in a very specific way, our clustered network model defines the appropriate ensemble to study the effect of clustering on network structure and function. In this direction, we are going to exploit our model to study the effect of clustering on one of the simplest network processes, the bond percolation problem [165].

Percolation theory emerged from the study of the movement of a liquid through a porous material. The question that scientist were interested in was: If a liquid is poured on top of some porous material; will the liquid be able to make its way from hole to hole and reach the bottom? This physical question is modelled mathematically as a three-dimensional lattice in which a fraction p of the edges are open (allowing the liquid through) and the rest are closed. Therefore, for a given value of p, what is the probability that a path exists from the top to the bottom?

In network theory in particular, the bond percolation problem studies the remaining structure of a network after a random removal of a fraction of its edges; providing an elegant theory of network robustness to random failures of their connections; a key issue for infrastructural and technological network design. Moreover, the bond percolation process can be mapped to one of the most popular epidemic spreading models, the Susceptible-Infected-Recovered model (SIR), so the bond percolation theory can also be used to predict the size of an outbreak of a given infected agent. Besides, percolation models have been used as a representation of resistor networks [7], forest fires [93], epidemics [128], biological evolution [97], and social influence [161].

5.1 The bond percolation problem

The bond percolation problem on networks can be stated as follows: given a network, we visit each edge and with probability p we preserve it or, which is the same, with probability $1-p$ we removed it. Then a primary question arises: which is the size of the remaining components as a function of the bond occupation probability p?

Bond Percolation is a classical problem that has attracted the attention of mathematicians and physicists for many years because it is one of the simplest models displaying a phase transition. Under this process, a connected system undergoes a continuous phase transition at a critical value p_c, known as the *percolation threshold*. Below p_c the network is made of a myriad of finite disconnected clusters. Above this critical value, a macroscopic cluster of the order of the size of the system, namely a giant component (GC), emerges, so the network becomes globally connected. Therefore, in practical applications, having an accurate prediction of the position of the percolation threshold is extremely important. For example, in infrastructural or technological networks this critical point separates the global functioning state from total collapse.

There have been many efforts to solve the bond percolation problem. In some situation an analytical solution is possible, so we can obtain an expression for the size of the GC as a function of the bond occupation probability p and the position of the critical point p_c. Examples in which we have an exact expression of p_c include the case of lattices of 1 and 2 dimensions and the Bethe lattice [165]. However, in many other systems we have no other option but to rely on numerical simulations.

5.2 Numerical simulations

5.2.1 Newman-Ziff algorithm

The most simple way to make numerical simulations to explore the bond percolation properties of networks is to directly visit each edge and remove it with probability $1-p$, and at the end, measure the size of the remaining clusters. However, this process is very expensive computationally. One would need to make independent simulations for all the different values of p, and for each value, make many simulations to get a proper average.

However, there is a much more clever way to do the simulations following the Newman-Ziff algorithm [137]. In each realization of this method, one starts from a configuration with no connections. We then sequentially add edges in random order and monitor the quantities that we are interested in, e.g. the size of the largest cluster in the network, G. We repeat the entire process to average

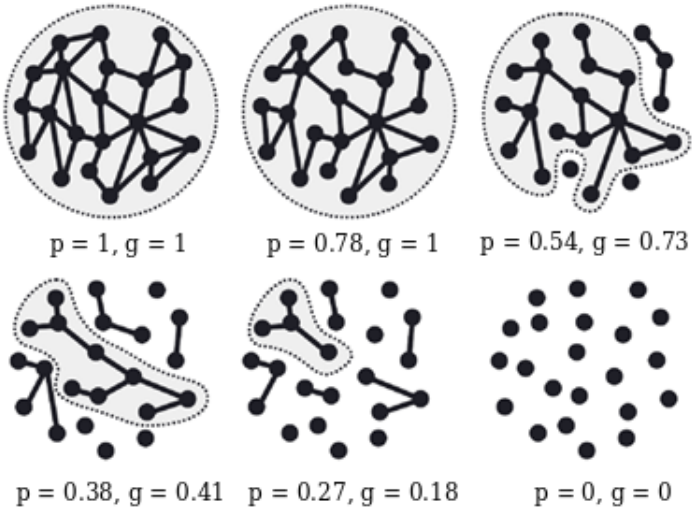


Figure 5.1: A simple example of a bond percolation process for different values of the bond occupation probability p and the relative size of the largest component g.

over many independent realizations and estimate the size of the largest component as a function of the number of edges added from the initial configuration, n. Then we can just find the value of G as a function of the bond occupation probability by convolution with the binomial distribution,

$$G(p) = \sum_n \binom{E}{n} p^n (1-p)^{E-n} G(n), \tag{5.1}$$

where E is the total number of edges of the network. This algorithm can find the value of a quantity or quantities over the entire range of p in time $O(N)$, a huge improvement over the simplest algorithm which performs as $O(N^2)$.

5.2.2 Percolation threshold and critical exponents

The critical point p_c determines the moment in which a macroscopic cluster of the order of the total system's size emerges, leading to a sudden increase of the size of the largest component of the network. Simulations are always performed on finite systems, so we cannot distinguish between a macroscopic cluster from a large one. Therefore, just from the observation of the size of the largest component we cannot determine the exact point in which a GC appears. Fortunately, continuous phase transitions are characterized by the divergence of the susceptibility of the order parameter at the critical point [165]. In our case the order

parameter is the size of the largest component so its susceptibility would simply correspond to its variance with a proper normalization [11, 114],

$$\chi_{st} \equiv \frac{\langle G^2 \rangle - \langle G \rangle^2}{N}. \tag{5.2}$$

In finite systems, the susceptibility χ will not diverge but will show a peak, and its position provides an estimate of the critical point p_c. Because the susceptibility should diverge in the thermodynamic limit, we expect that the height of the peak of χ should increase with the system size. In particular, according to the finite size scaling assumption, the susceptibility χ_{st} should behave as

$$\chi_{st}(N,p) = N^{\gamma/\nu} F(|p - p_c| N^{\frac{1}{\nu}}). \tag{5.3}$$

where $F(x)$ is scaling function that behaves as $x^{-\gamma}$ far from the critical point, $x >> 1$, and is constant close to the critical point, $x << 1$ [165]. Therefore, the height of maximum of the susceptibility should depend on the system's size as $\chi_{st}(p) \propto N^{\gamma/\nu}$. At the same time, the position of the maximum of the susceptibility, p_{max}, moves towards the theoretical value p_c as $p_{max} - p_c \propto N^{-1/\nu}$.

The same scaling assumption is valid for the size of the largest connected component. Close to the critical point the GC should depend on the system size as $G \propto N^{-\beta/\nu}$.

The exponents γ, β and ν are critical exponents[1]. In percolation theory, the critical exponents are a set of universal parameters that depend on the dimensionality of the network but not on microscopic details of the system. Therefore, the measurement of the critical exponents characterize a phase transition allowing for a proper classification. For instance, all regular lattices with a higher dimension than 6, such as a Cayley tree, have the same mean-field values $\beta = \gamma = 1$ and $\nu = 3$ [165] (See Appendix A.2 for more details). This regular mean-field result is not always valid, however, for scale-free networks [47] .

In this thesis, in order to measure the critical point we use the susceptibility of the giant component but with a different normalization,

$$\chi \equiv \frac{\langle G^2 \rangle - \langle G \rangle^2}{\langle G \rangle}. \tag{5.4}$$

The advantage of using Eq. (5.4) instead of Eq. (5.2) is mainly numerical for measuring the critical exponents. For a finite system of size N, the peak of the standard susceptibility near the critical point behaves as $\chi_{st}^{max} \sim N^{\gamma/\nu}$ and our version of the susceptibility χ diverges as $\chi \sim N^{\gamma'/\nu}$, where $\gamma' = \gamma + \beta$. This means that $\gamma' > \gamma$ and, thus, it is easier to measure in numerical simulations. This method is assumed to give the most accurate measure of the bond percolation threshold [148].

[1] Do not confuse the critical exponent γ with the γ used in the exponent of the scale-free degree distributions $P(k) \sim k^{-\gamma}$.

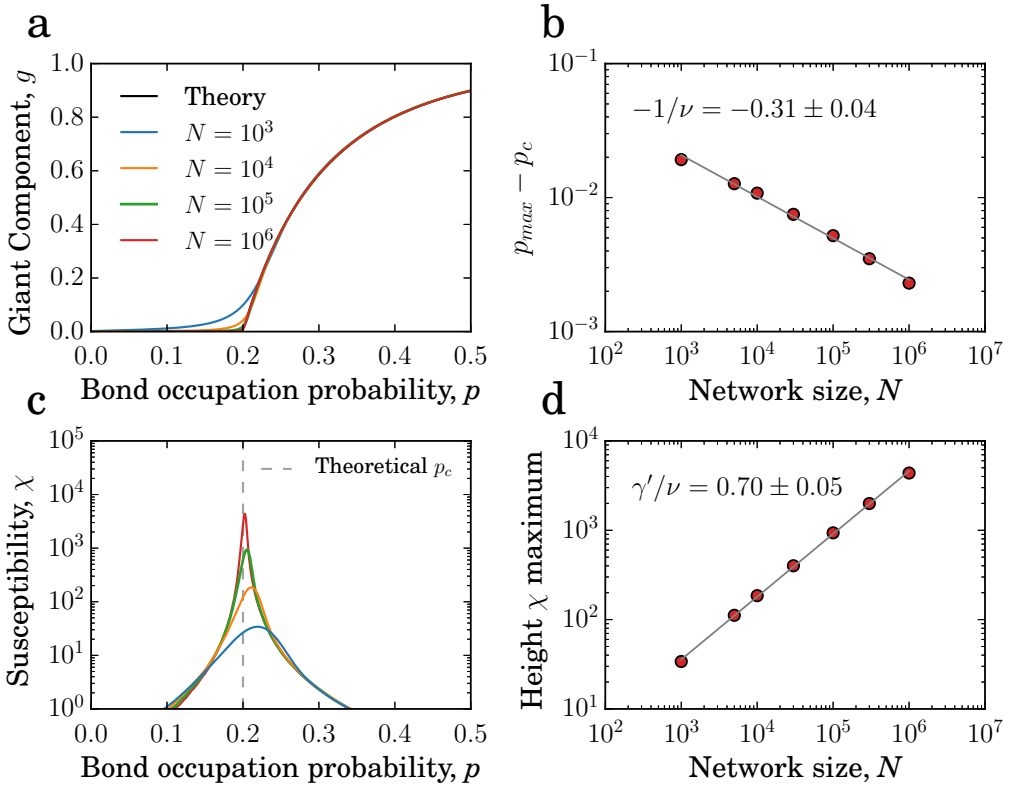


Figure 5.2: Bond percolation simulations for Erdős-Rényi networks with an average degree $\langle k \rangle = 5$ and different network sizes. **a)** Relative size of the largest connected component g as a function of the bond occupation probability p. **c)** Susceptibility χ as a function of the bond occupation probability p. **b)** Position p_{max} minus the theoretical value of the percolation threshold $p_c = 0.2$ as functions of the network size N. **d)** Height χ_{max} of the peak of χ as functions of the network size N. The straight lines are power-law fits, and **b** and **d** show the measured values of the critical exponents.

Figure 5.2 shows bond percolation simulations on Erdős-Rényi networks with an average degree $\langle k \rangle = 5$ for different system sizes. In figure 5.2 **a**, we can observe how the size of the largest connected component suddenly increases at a value close to $p = 0.2$. The finite size scaling of the susceptibility in figure 5.2 **c** shows that the susceptibility of the giant component χ has a peak close to the critical point that increases and gets closer to the theoretical value $p_c = 0.2$ as we increase the system size [37]. Figures 5.2 **b** and **d** use these phenomena to calculate the critical exponents of the phase transition. The values we obtain agree with the mean field theoretical values [165](See appendix A.2 for more detailed explanation of the mean field values).

5.3 Bond percolation on random networks

Although the numerical simulations provide the more accurate prediction of the size of the largest connected component as a function of the bond occupation probability p and the position of the critical point, they do not give any information on the relation between network structure and its percolation properties. Therefore, analytic solutions of the bond percolation problem are useful to understand the role that each topological property has on network robustness to the random failure of connections.

One example of our interest, in which an analytic solution is possible, is that of random network with a given degree distribution, the so called configuration model [37]. In this case, we can solve the problem giving a lot of insights on how the degree distribution affects the bond percolation properties of networks.

Consider a network generated using the configuration model with degree distribution $P(k)$ in which only a fraction p of its edges are preserved and the rest are removed. We define u as the average probability that a node is not connected to the giant component (GC), via a particular neighbour. Then, the fraction of nodes that belong to the GC, g, is equal to the probability that a randomly chosen node belongs to the GC, or one minus the probability that a random chosen node does not belong to the GC, so

$$g = 1 - \sum_k P(k) \sum_n^k \binom{k}{n} p^n (1-p)^{k-n} u^n. \tag{5.5}$$

Note that we used the binomial distribution to calculate the probability that, after the percolation process on the previous k connections that a node had, n of them are preserved and $k-n$ are removed. We then expressed the conditional probability that a node does not belong to the GC as the product of the independent probabilities that it does not belong to the GC via each of its n remaining neighbours, which is u, so we get the term u^n. This last step assumes that none

of the neighbours of the node we are considering are connected by any path that do not go through the node itself, that is only possible in a network that has no close loops, namely a tree network. The type of random networks we are considering are not exactly trees but, their small presence of short loops makes this assumption valid, at least locally.

Applying the binomial theorem to equation 5.5 we get

$$g = 1 - \sum_k P(k)(1 - p + pu)^k. \tag{5.6}$$

Here is important to point out that $u = 1$ gives $g = 0$ so there is not giant cluster, and $u = 0$ means that the giant cluster is the whole network. However, we still need to calculate an expression for u. The probability that a node i does not belong to the GC through one of its neighbour is equal to the probability that this neighbour, with degree k, does not belong to the GC through any of its $k-1$ other neighbours rather than node i. Using the same technique that we used in Eq. 5.5 we can express u as

$$u = \sum_k P(k|k') \sum_n^{k-1} \binom{k-1}{n} p^n (1-p)^{k-1-n} u^n. \tag{5.7}$$

where $P(k|k')$ is the probability that the node i of degree k' is connected to a node of degree k which in our case, because the configuration model is uncorrelated in terms of degrees, takes the form $kP(k)/\langle k \rangle$. Then, Eq. 5.7 leads to

$$u = \sum_k \frac{kP(k)}{\langle k \rangle} (1 - p + pu)^{k-1}. \tag{5.8}$$

Equations 5.6 and 5.8 give us a complete solution for the size of the giant cluster in our network. A numerical solution of these equations would gives us the theoretical value of the size of the GC as a function of the bond occupation probability. However, having a theoretical curve given by a numerical solution of these equations is not giving much more information than one would get from direct simulations. To get valuable knowledge, we need to solve these equations in closed form and get an expression of the GC or the percolation threshold as a function of some properties of the degree distribution.

In practice it is often not possible to solve Eq. 5.8 in a closed form, but there is an elegant graphical representation of the solution as follows. Because all the variables of Eq. 5.8 are definite positive, and p and u are smaller than 1, we know that the functions on both sides of the equality are increasing functions with u. Besides, we know that both functions cross at the trivial solution $u = 1$, which implies there is no GC. Only if there is a non-trivial solution there can be a giant cluster. Because the right-hand side of Eq. 5.8 is greater than 0 for $u = 0$, we know

that in order to have one non trivial solution the derivative of the right hand side at $u = 1$ must be larger than the derivative of the left-hand side at the same point, which is equal to 1. Therefore, the bond percolation threshold p_c is given by the equation

$$\left[\frac{\mathrm{d}}{\mathrm{d}u} \sum_k \frac{kP(k)}{\langle k \rangle} (1 - p_c + p_c u)^{k-1} \right]_{u=1} = 1. \tag{5.9}$$

If we apply the derivative operator we obtain

$$\sum_k \frac{k(k-1)P(k)}{\langle k \rangle} p_c = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} p_c = 1, \tag{5.10}$$

which leads to the expression for the bond percolation threshold

$$p_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}. \tag{5.11}$$

Equation 5.11 describes the exact impact of the degree distribution on the percolation threshold. From this expression we see that more heterogeneous networks have a smaller percolation threshold. Interestingly, the common scale-free distribution present in many real networks, with a power law behaviour $P(k) \sim k^{-\gamma}$, with exponent $\gamma < 3$, have a infinite second moment, leading to a vanishing percolation threshold. This fact implies that heterogeneity in the connectivity of the nodes present in real networks makes them very resilient to random failure of their connections.

Figure 5.2 **a** compares the results of bond percolation simulations on an Erdös-Rényi graph and its theoretical curve given by Eqs. 5.8 and 5.6. As we can see, both curves fit very well which implies that the Tree-like assumption that we used in our calculations holds reasonably well on the classical random graph model. Moreover, the value of the percolation threshold given by Eq. 5.11 is $p_c = 0.2$ which is very close to the position of the maximum of the susceptibility of the giant component.

5.4 Bond percolation and epidemics

Computer viruses, technological innovation, rumours, beliefs, or viral marketing campaigns spread on the population through social contacts the same way an infectious disease does. Although every process has its own particularity, all of them can be modelled using epidemiological models, in which infected individuals have a certain probability to infect their contacts.

Traditional epidemiological models make use of the so called "fully mixed" approximation in which all individuals are connected among themselves [5, 8, 94]. However, in the real world the potential contacts of individuals are restricted

to finite number of individuals that can represent acquaintances, neighbours, co-workers, and so forth. Because not every body has the same number of contacts, the set of potential contacts that each person have can be represented as a network, whose structure can have a strong effect on the epidemic dynamics. Therefore, the study of the effect of network topology on different epidemiological models has been an important breakthrough on the understanding of disease dynamics.

One of the most popular epidemiological model is the Susceptible-Infected-Recovered (SIR) model [90, 125]. This model is one of the most fundamental epidemic models that can be applied to many common bacterial o viral infections that confer upon their host a certain immunity to catch the disease again, or are very deadly. In this model every individual can be in three possible states. An individual in the susceptible state is someone that does not have the disease but can be infected if he or she comes in contact with someone who has the disease. An infected individual is someone that has the disease and can propagate it coming in contact with susceptible individuals. A recovered individual is someone who either has become immune to the disease or he or she has died. From a mathematical point of view it does not matter if the recovered individual is immune or dead, because in both situations this individual cannot become infected again or infect any other individual.

The dynamics of the SIR model is as follows: within a period $\delta\tau$, any infected agent can infect any of its susceptible neighbours with probability λ or recover with probability β. Because in the SIR model infected individuals remain infected a finite amount of time and then become recovered, it is possible that they can recover before they have been able to spread the disease. Given an edge between an infected node and a susceptible one, the probability that the infected agent does not infect the susceptible node after a period of time τ is $(1-\lambda\delta\tau)^{\tau/\delta\tau}$, which in the continuous time limit we obtain

$$\lim_{\delta\tau\to 0}(1-\lambda\delta\tau)^{\tau/\delta\tau} = e^{-\lambda\tau}. \tag{5.12}$$

Then, the probability $P(\tau)\delta\tau$ that the susceptible individual remains healthy this long and then becomes infected in the interval between τ and $\tau+\delta\tau$ is

$$P(\tau)\delta\tau = \lambda e^{-\lambda\tau}, \tag{5.13}$$

Similarly to Eq. 5.12, the probability that an infected agent does not recover after a period of time τ is equal to $e^{-\beta\tau}$. Therefore, the probability ϕ that the disease is transmitted through one edge before the infected node becomes recovered is given by

$$\phi = \int_0^\infty \lambda e^{-\lambda\tau} e^{-\beta\tau} \mathrm{d}\tau = \frac{\lambda}{\lambda+\beta} = \frac{R}{R+1} \tag{5.14}$$

where R is the basic reproduction number, the key parameter of a disease given by the ration between the transmissibility λ and the mortality/latency β parameters, $R \equiv \lambda/\beta$.

Equation 5.14 implies that every edge, with probability ϕ, will be able to transmit the disease if it reaches one of its nodes at the end. Instead, with probability $1-\phi$, an edge will never propagate the disease even if a node at the end becomes infected. Thus, if we are only interested in the late time state of an SIR disease, we could first remove randomly a fraction $1-\phi$ of the edges and assume that the remaining edges will always propagate the disease. This procedure is equivalent to a bond percolation process in which the probability ϕ given by the Eq. 5.14 plays the role of the bond occupation probability p and the cluster's size of the remaining network in which the outbreak started is the final prevalence of the disease.

Therefore, the late time state of the SIR model can be mapped to a bond percolation problem. Hence, the bond percolation threshold p_c corresponds precisely to the epidemic threshold, R_c, in which the SIR model has a phase transition that separates the endemics state, in which a macroscopic portion of the population can become infected, from the healthy state, in which the disease will die out before becoming pandemic. Actually, using the previous results for the percolation threshold in Eq. 5.11 combined with Eq. 5.14, we can find an expression for the epidemic threshold of a network generated by the configuration model in terms of the degree distribution:

$$R_c = \frac{\langle k \rangle}{\langle k^2 \rangle - 2\langle k \rangle}. \tag{5.15}$$

However, this mapping between bond percolation and the SIR model is not completely exact. Using the bond percolation problem we get accurate predictions for the mean outbreak size below the epidemic threshold, the same epidemic threshold, and the same final size of an epidemic. However, the bond percolation model fails to predict the correct outbreak size distribution and probability of an epidemic. The contact rate pairs for all edges incident to a susceptible node are independent. However, transmission events from the same infected node are marginally dependent, since they depend on the recovery of the same node, unless the recovery time distribution is a delta function [102, 121]. Nevertheless, the inaccuracies of the bond percolation mapping of the SIR model can be fixed applying the bond percolation process on a semi-directed random network of the type introduced in [23].

Nevertheless, the mapping between bond percolation and epidemics is a powerful fact that allows us to apply many results on the bond percolation problem to the spread of diseases. Therefore, the direct relation that the bond percolation process has with network structure and robustness, but also with network

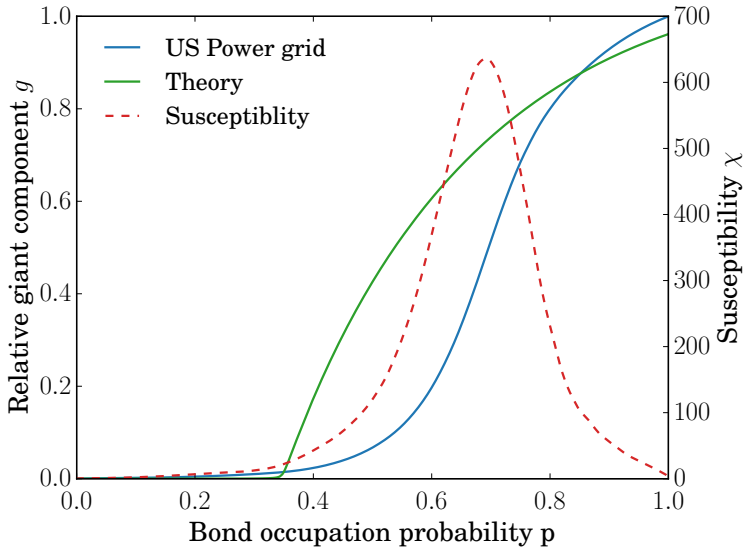


Figure 5.3: Comparison of the relative size of the giant component g (left axis) as a function of the bond occupation probability p from simulations of the western union states of US power grid. The theoretical curve is given by equations 5.6 and 5.8. On the right axis we plotted the susceptibility of the giant component for the real network (red dashed line).

dynamics, makes this classic problem the perfect framework to study the role that clustering plays on network structure and function.

5.5 Bond percolation on real networks

As in section 5.3, all the analytical results for the bond percolation problem assume that the networks are locally tree-like [33, 37, 46, 99]. However, the high level of clustering in empirical networks cast doubts on the validity of the Tree-like assumption in real cases [148]. In some situations these theories still perform well [117], but in a large number of real cases the inaccuracies of these theories are shown to be only caused by the presence of short loops [71]. For instance, Fig. 5.3 compares bond percolation simulations on the Western union states of United States power grid with the theoretical curves given by Eqs. 5.6 and 5.8. As we can clearly see, the theoretical curve deviates substantially from numerical simulations. Since the theories we have at hand are very accurate for locally tree-like networks, it becomes evident that clustering is an important missing piece of the puzzle.

In an effort to overcome these problems, a new class of network models

had been proposed in which the resulting clustered graph is embedded in another graph that is locally tree-like, thus allowing for an analytical treatment. [83, 84, 85, 86, 98, 122, 134, 167]. However, as we have seen in the previous chapter, triangles generated by these models are arranged in a very specific way, with strong correlations between the properties of adjacent edges, not present in real networks. Nevertheless, our maximally random clustered network model defined in section 2.5 does reproduce the global organization of triangles in real networks. Therefore, we consider that our model defines an appropriate framework to study how clustering affects the bond percolation properties of networks.

Because all the mathematical tools to solve the bond percolation problem at some point rely on the tree-like assumption, the possibility to have analytical solutions that includes clustering does not look promising. Therefore, in the next chapter we are going to exploit our clustered network model to make an empirical study on how clustering effects the position of the bond percolation threshold.

CHAPTER 6

Bond percolation on clustered networks

6.1 Bond percolation on clustered networks

Percolation theory has played a prominent role in understanding the anomalous behaviors observed in complex networks and, in most cases, is the common underlying principle behind these behaviors. Interestingly, the interplay between a complex network topology and different percolation mechanisms leads to phenomena that have not previously been observed in statistical physics, including a lack of percolation thresholds in scale-free networks with a degree distribution of the form $P(k) \sim k^{-\gamma}$ for $\gamma < 3$ [18, 24, 29, 41, 112, 144], anomalous infinite-order percolation transitions in non-equilibrium growing random networks [38, 64], or cascading processes in interdependent networks [14, 35, 164]. However, these phenomena have already been observed on random graphs with given degree distributions. Random graphs of this type are locally tree-like, that is, the number of triangles, and thus the clustering coefficient, can be neglected in the thermodynamic limit. However, as we have seen, the strong presence of triangles is, along with the small-world effect and heterogeneity of the degree distribution, a common and distinctive topological property of many real complex networked systems. While clustering is not a necessary condition for the emergence of any of these phenomena, the effects of clustering on the percolation properties of a network are unknown.

Percolation in clustered networks has been widely studied [84, 86, 104, 122, 132, 134, 154]. However, previous reports differ concerning the position of the percolation threshold. Some studies report that clustered networks have a larger percolation threshold than do unclustered networks due to redundant edges in triangles that cannot be used to connect to the giant component (GC) [86, 104, 122, 134]. Other studies report that strongly clustered networks are more resilient due to the existence of a core that is extremely difficult to break [84, 132, 154]. In fact, as we shall demonstrate, both arguments are correct.

Here, we show that strong clustering induces a core-periphery organization in the network [54] that gives rise to a new phenomenon, namely, a “double

percolation" transition, in which the core and periphery percolate at different points. This behaviour is in stark contrast to the modern theory of continuous phase transitions, which forbids the possibility of breaking the same symmetry at two different values of the control parameter. Multiple percolation transitions have recently been reported in [19, 42, 43, 129]. However, in each of these cases, anomalous percolation arises as a consequence of either complex percolation protocols [42, 43, 129] or the interdependence between different networks [19], and it is never associated with the same symmetry breaking. Instead, our results are obtained with the simplest percolation mechanism, bond percolation with bond occupation probability p, which indicates that this double percolation transition is exclusively induced by a particular organization of the network topology.

6.2 Random graphs with a given clustering spectrum

To study empirically the effects of varying one network property (e.g. clustering), one would ideally like to generate multiple networks with all properties identical, except the property of interest. However this task is not easy to put it in practice because network properties may constrain each other, or not be independent. Here, in order to study the effect of clustering on the bond percolation properties of networks we will use our maximally random clustered networks to compare networks with the same degree distribution and degree-degree correlations and different clustering spectrum.

Our results will be only strictly valid for our maximally random networks. However, in chapter 4 we showed that our maximally random clustered network model is the one that better reproduced the global organization of real complex networks. Outperforming previous clustered network models. Besides, a preliminary analysis in Ref. [154] shows that the percolation properties depend on two network features, the joint degree distribution $P(k,k')$, and the shape of the clustering spectrum $\bar{c}(k)$. Moreover, Ref [140] shows that long-range properties of real complex networks are very well reproduced by maximally random networks with the same degree distribution, degree-degree correlation and clustering spectrum. Therefore, we expect that our result here can also be applied to real networks.

To generate scale-free random graphs with a given clustering spectrum $\bar{c}(k)$ and fixed degree-degree correlations we use the model developed in section 2.5. Therefore, we first generate a degree sequence according to a desired degree distribution. Then, from this degree sequence, we generate a random network using the configuration model. Finally, we add the desired level of clustering using the rewire Metropolis-Hastings algorithm together with the annealed procedure.

Our model allows to give any desired clustering spectrum. Since the local clustering coefficient of a node is the probability that two random neighbours are connected, we decided to give the same probability to all nodes. This choice corresponds to a constant or flat clustering spectrum. This case is not the same as fixing the average clustering coefficient because, in our case, we enforce nodes of any degree to have the same local clustering (see Fig. 2.2 for a comparison). In any case, due to structural constraints, for very strong clustering it is not possible to keep $\bar{c}(k)$ constant for very large values of k. In this case, the algorithm generates the maximum possible clustering [155].

For each network, we perform bond percolation 10^4 times using the Newman-Ziff algorithm [137] and measure the average relative size of the largest (giant) connected component, $g \equiv \langle G \rangle / N$, and its fluctuations, *i.e.*, the susceptibility $\chi = \left[\langle G^2 \rangle - \langle G \rangle^2 \right] / \langle G \rangle$. These results are then averaged over 100 network realizations. In finite systems, a peak in the susceptibility χ indicates the presence of a continuous phase transition, and its position provides an estimate of the percolation threshold.

6.3 Weakly heterogeneous networks

We first studied the scale-free networks, which have a degree distribution that follows a power law $P(k) \sim k^{-\gamma}$, in the weakly heterogeneous regime ($\gamma \gg 3$). Figure 6.1 show the results of the bond percolation properties of networks with the same degree distribution, for $\gamma = 3.5$, and degree-degree correlations but different levels of clustering. All networks have a unique and well defined peak in the susceptibility χ, and an increase of the clustering moves the peak to higher values of p. So in this situation, according to [86, 104, 122, 134], triangles are redundant edges that can not be used to connected the GC together. We obtained the same results for other networks with larger values of γ or with a Poisson degree distribution which are even less heterogeneous than the one just reported. Hence clustering decreases the GC and increases the percolation threshold of weakly heterogeneous networks.

6.4 Heterogeneous networks

However, most of real networks are heterogeneous so the most interesting case corresponds to heterogeneous networks, typically with $\gamma < 3.5$. In particular we focus on the case of $\gamma = 3.1$ and a constant clustering spectrum. This value of γ generates scale-free heterogeneous networks but with a finite second moment, which allows us to clearly isolate the new phenomenon. The results for $\gamma \leq 3$ are qualitatively similar.

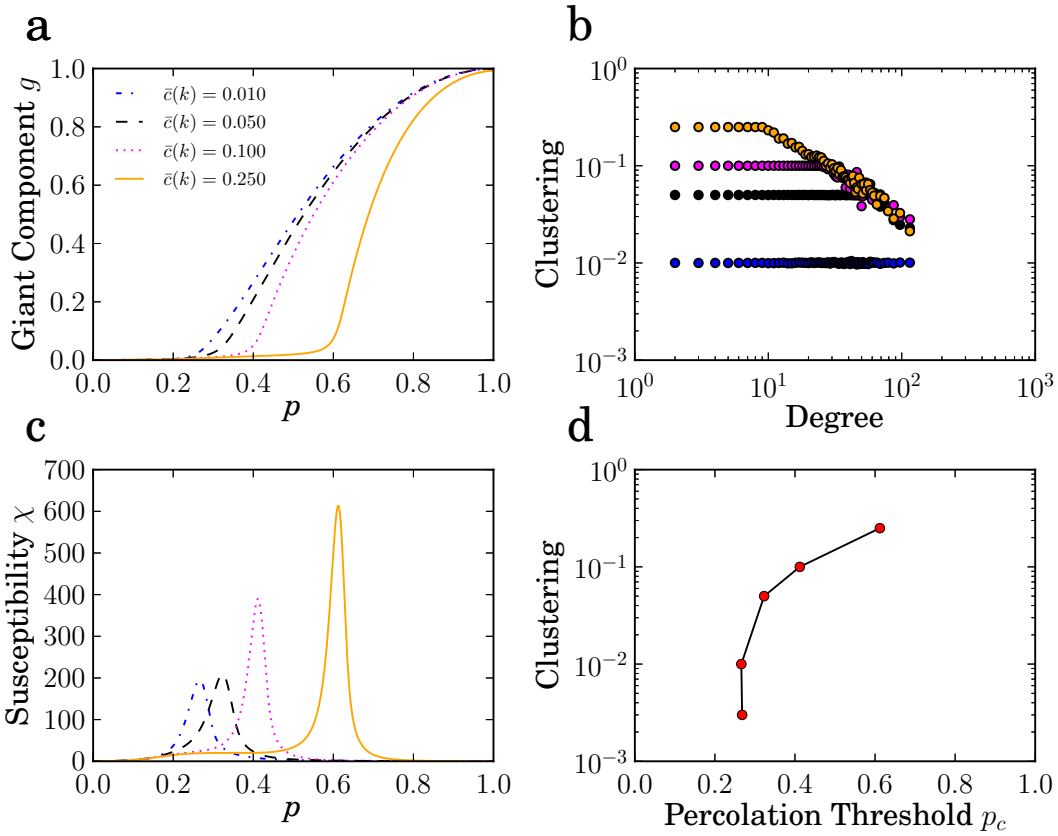


Figure 6.1: Bond percolation simulations for networks of 10.000 nodes with a power law degree distribution with $\gamma = 3.5$ and different levels of clustering. **a)** Relative size of the largest connected component g as a function of the bond occupation probability p. **b)** Degree-dependent clustering coefficient $\bar{c}(k)$. **c)** Susceptibility χ as a function of bond occupation probability p. **d)** Percolation threshold (p_{max}) as a function of the level of clustering.

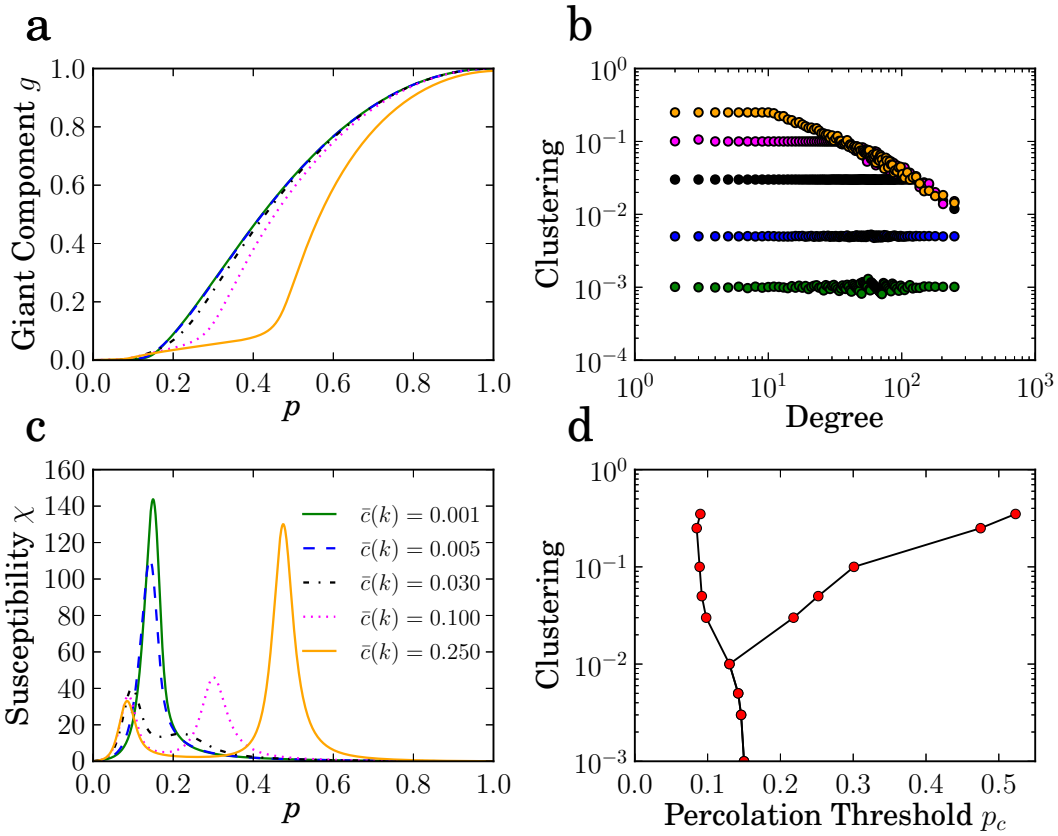


Figure 6.2: Bond percolation simulations for networks of $N = 5 \times 10^4$ nodes with a power law degree distribution, $\gamma = 3.1$, and different levels of clustering. **a)** Relative size of the largest connected component g as a function of the bond occupation probability p. **b)** Degree-dependent clustering coefficient $\bar{c}(k)$. **c)** Susceptibility χ as a function of the bond occupation probability p. **d)** Percolation threshold (p_{max}) as a function of the level of clustering.

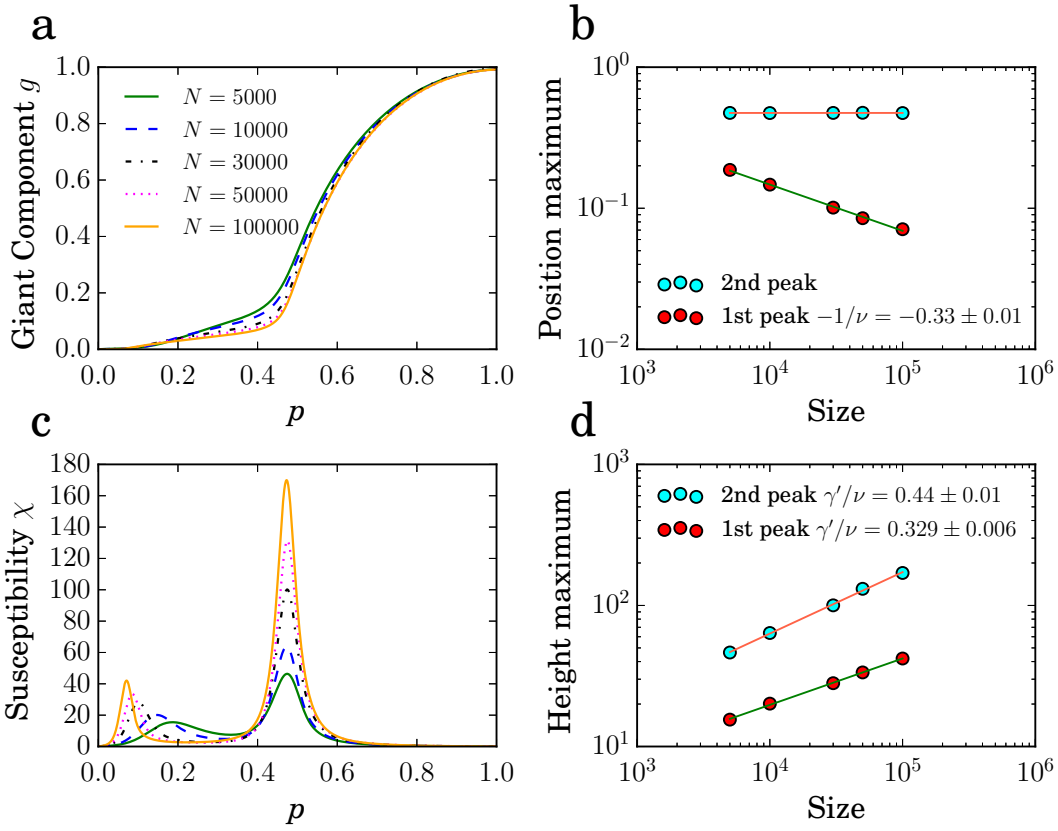


Figure 6.3: Bond percolation simulations for networks with a power law degree distribution with $\gamma = 3.1$, target clustering spectrum $\bar{c}(k) = 0.25$, and different network sizes. **a)** Relative size of the largest connected component as a function of the bond occupation probability p. **c)** Susceptibility χ as a function of the bond occupation probability p. **b** and **d:** Position p_{max} and height χ_{max} of the two peaks of χ as functions of the network size N respectively. The straight lines are power-law fits and the measured values of the critical exponents are shown.

Figure 6.2 compares the percolation properties of networks with identical degree sequence ($\gamma = 3.1$) and degree-degree correlations but with different levels of clustering. Plots **c** and **d** in Fig. 6.2 show new and surprising results. For low levels of clustering, there is a unique and well-defined peak in χ, but increasing clustering gives rise to the emergence of a secondary peak at higher values of p. This result suggests the presence of a double phase transition, in which two different parts of the network percolate at different times.

To confirm this possibility, we perform finite size scaling on networks with a target clustering spectrum of $c(k) = 0.25$ and different system sizes, ranging from $N = 5 \times 10^3$ to $N = 5 \times 10^5$. Plot **d** in Fig. 6.3 shows that the susceptibility exhibits two peaks whose maxima χ_{max} diverge as power laws, $\chi_{max}(N) \sim N^{\gamma'/\nu 1}$. The position of the first peak also approaches zero as a power law $p_{max}(N) \sim N^{-1/\nu}$, as shown in Fig. 6.3 **b**, which suggests that even if the network has bounded fluctuations, $\langle k^2 \rangle < \infty$, it is always percolated in the thermodynamic limit. In contrast, the position of the second peak is nearly constant in the range of sizes we have considered. The divergence of the two peaks in the susceptibility strongly suggests that we are indeed observing two different continuous phase transitions. The first transition is between non-percolated/percolated phases, and the second transition is between two percolated phases with very different internal organizations.

6.5 The clustering m-core decomposition

To understand the effect of clustering on the global structure of networks, we use the clustering m-core decomposition developed in section 4.3. This process is based on the concept of edge multiplicity m, which is defined as the number of triangles passing through an edge. We further define the m-core as the maximal subgraph whose edges all have at least multiplicity m within it. By increasing m from 0 to m_{max}, we define a set of nested sub-graphs that we call the m-core decomposition of the network. This decomposition can be represented as a branching process that encodes the fragmentation of m-cores into disconnected components as m is increased. The tree-like structure of this process provides information regarding the global organization of clustering in networks. To visualize this process, we use the LaNet-vi 3.0 tool developed in section 4.4. Figure 6.4 shows the m-core decomposition of three networks with $N = 5 \times 10^4$ nodes, the same degree sequence (with $\gamma = 3.1$) and degree-degree

[1] Note that we evaluate the size dependence as a function of the total number of nodes N and not as the one dimensional length L. This implies that the finite size critical exponent ν that we use in this paper already includes the dimensionality of the system d, that is, our exponent ν is equivalent to $d\nu$ in [165].

correlations, and different levels of clustering. For low levels of clustering, the $m1$-core is very small, and thus, the m-core structure is almost non-existent. As clustering increases, m-cores begin to develop new layers and m_{max} increases. For instance, for $\bar{c}(k) = 0.25$ (Fig. 6.4 **c**), after the recursive removal of all links that do not participate in triangles, we obtain the $m1$-core, which is composed of a large connected cluster with a well-developed internal structure – a core in the center of the figure – and a large number of small disconnected components – a periphery. This result indicates that even if the network is connected, by iteratively removing all edges with multiplicities of zero, we are left with a small but well-connected subgraph and the reminder of the network is fragmented. Drastic topological transitions induced by clustering have been also reported in the Strauss model and its generalizations [73, 143, 166].

In case of weakly heterogeneous networks $\gamma > 3.5$. The m-core visualization of networks shows that there is no well entangled core even for large levels of clustering. In this case there are no enough hubs to create a robust core.

6.6 Identification of the core

The aforementioned result suggests that the two peaks in the susceptibility of the GC in heterogeneous networks could be related to this core-periphery organization. Both parts would percolate at different times, first the core and then the periphery, and hence have their own percolation thresholds. To test this hypothesis, we have to perform independent bond percolation simulations on the core and the periphery. In order to identify which nodes belong to the core and which to the periphery we perform a bond percolation simulation on a network of 50000 nodes $\gamma = 3.1$ and $c(k) = 0.25$. We first delete all edges and then we add the edges one by one randomly. Once we added a 20% of the total number of edges ($p = 0.2$ that lays between the two percolation thresholds) the giant component (GC) defines a subgraph that we identify with the core (red nodes in Fig 6.6), that roughly corresponds to the core observed in Fig. 6.4 **c**. If, in the same simulation, we keep adding edges we will observe another phase transition where the periphery percolates at $p = 0.5$. However the periphery has percolated regardless of the core. This can be observe if we subtract the nodes that belong to the core and see that largest component that remains is still a macroscopic component (blue nodes at Fig. 6.6), and only few nodes leave the GC (green nodes in Fig. 6.6).

Once the core and periphery are isolated, we perform bond percolation on both components independently and compare the results with the original network. Figure 6.5 shows that the core percolates precisely at the point where the first peak appears in the original network, whereas the periphery percolates at

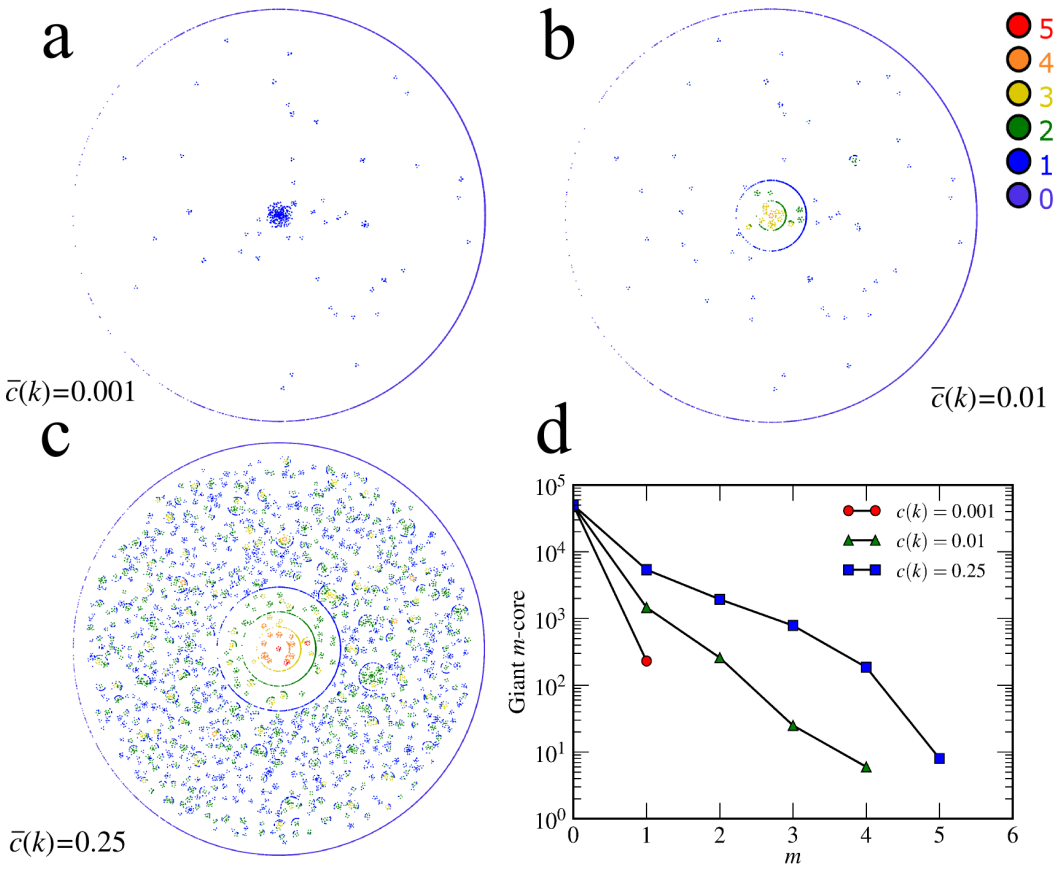


Figure 6.4: **a–c:** clustering m-core decomposition of three different networks with $N = 5 \times 10^4$, $\gamma = 3.1$, and different levels of clustering, $\bar{c}(k) = 0.001, 0.1, 0.25$. The color code of a node represents its m-coreness. For instance, nodes coloured violet belong to the $m0$-core but not to the $m1$-core and are said to have m-coreness of zero. The blue coloured nodes belong to the $m1$-core but not to the $m2$-core and have m-coreness of 1, etc. The visual representation is as follows. The outermost circle and its contents represent the $m0$-core and therefore the entire network. If we recursively remove all edges of multiplicity 0, we obtain the $m1$-core subgraph, which is contained within the $m0$-core. Nodes with no remaining connections do not belong to the $m1$-core, have m-coreness of 0, and are located at the perimeter of the outermost circle. If the $m1$-core is fragmented into different disconnected components, they are represented as non-overlapping circles within the outermost one and with nodes of m-coreness of 1 located in their perimeters (see, for instance, panels b and c). The same process is repeated for each disconnected $m1$-core, which will contain a subset of the $m2$-core, and so on. Links between nodes are not depicted for clarity. **d)** The size of the giant m-core as a function of m for the networks shown in panels **a–c**.

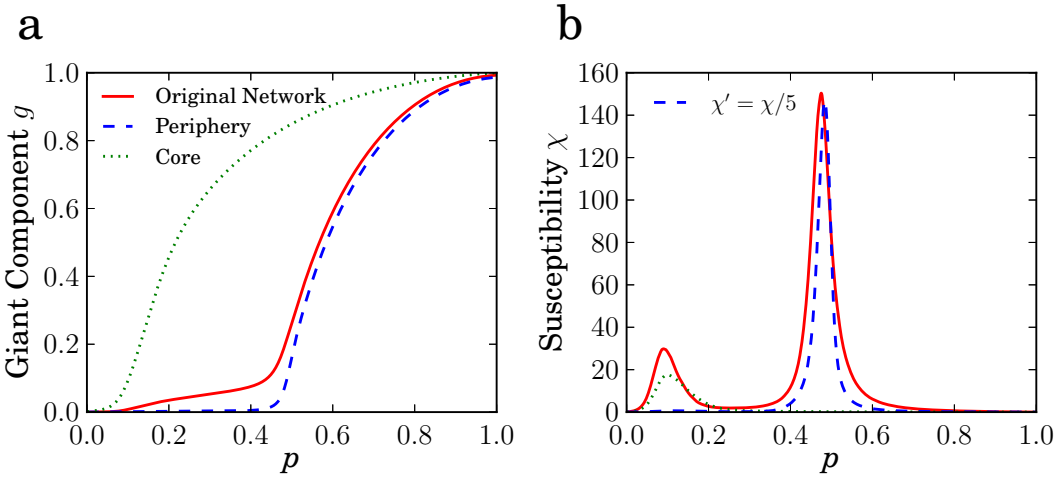


Figure 6.5: Bond percolation simulations of the core and periphery of a network with $N = 5 \times 10^4$, $\gamma = 3.1$, and target clustering spectrum $\bar{c}(k) = 0.25$. The bond occupation probability to separate the core is $p = 0.2$. The susceptibility curve of the periphery (dashed blue line) has been divided by 5 for ease of comparison.

the second peak.

6.7 The core-periphery random graph: a simple model showing a double percolation transition

The modern theory of continuous phase transitions states that, in a connected system, it is not possible to break the same symmetry at two different values of the control parameter. In our context, this statement implies that it is not possible to have two genuine percolation transitions at two different values of p. It is then unclear whether the second peak observed in our simulations corresponds to a real percolation transition or to a smeared transition, with the percolated core acting as an effective external field that provides connectivity among nodes in the periphery.

Unfortunately, strongly clustered networks cannot be studied analytically. However, we can devise a system with a core-periphery organization similar to that induced by strong clustering. Let us consider two interconnected random graphs a and b with average degrees $\bar{k}_{aa}$ and $\bar{k}_{bb}$, respectively. The relative size is $r = N_a/N_b$ and the average number of connections of a node in a to nodes in b (and vice versa) are $\bar{k}_{ab}$ and $\bar{k}_{ba} = r\,\bar{k}_{ab}$. Each node has connections to both networks and therefore its degree can be represented as a vector $\vec{k} = (k_a, k_b)$. Hence

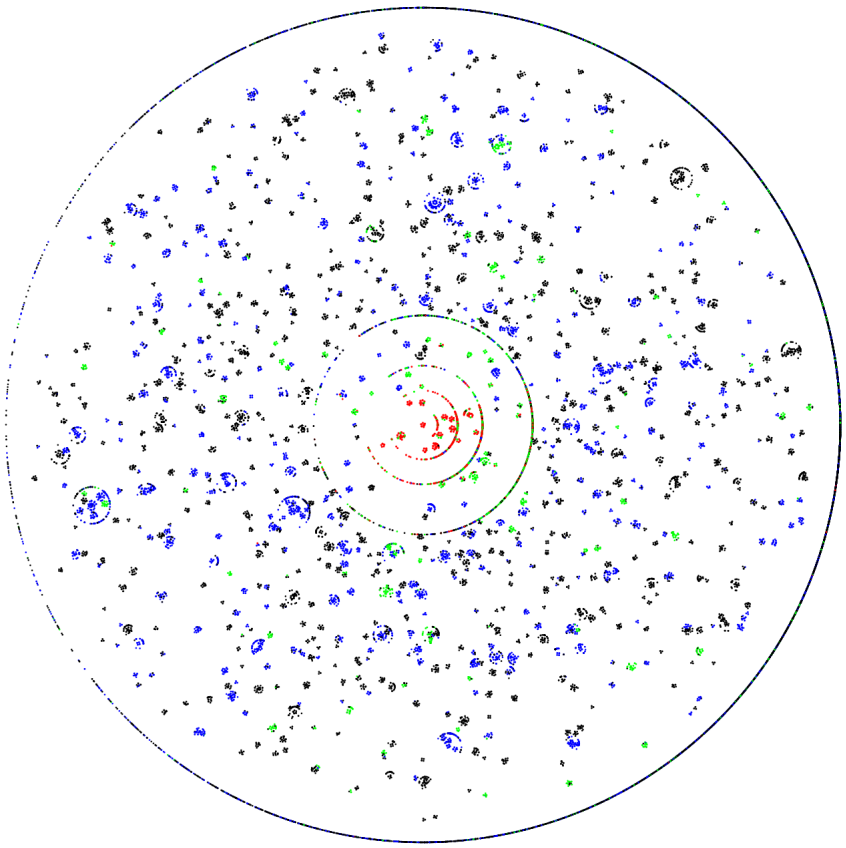


Figure 6.6: A network of 50.000 nodes, with a power law degree distribution with $\gamma = 3.1$ and a clustering spectrum $\bar{c}(k) = 0.25$. The nodes are distributed according to its m-core decomposition. Red nodes (1811) are the core they because belong to the Giant component once we perform a bond percolation with $p = 0.2$ (between the two percolation thresholds). Blue and green nodes are peripheral nodes that belong to the giant component at $p = 0.5$ (just after the second percolation threshold). Once we subtract the core, blue nodes (10408) still remain in the GC meanwhile green nodes (4271) belong to small components. Black nodes (33510) never belong to the GC.

$P_a(\vec{k})$ is the probability of a node of the network a to have degree $\vec{k}$ and $P_{ab}(\vec{k}'|\vec{k})$ is the probability that a node of a with degree $\vec{k}$ is connected to a node of b with degree $\vec{k}'$. The relative size of the giant component of the combined network is

$$g(p) = \frac{r}{1+r} g_a(p) + \frac{1}{1+r} g_b(p). \tag{6.1}$$

Where g_a is the probability that a node of a belongs to the giant component, or 1 minus the probability that it belongs to a finite cluster, that is, $g_a = 1 - \sum_{s=0}^{\infty} Q_a(s)$, where $Q_a(s)$ is the probability that a randomly chosen node from network a belongs to a cluster of size s.

In heterogeneous networks, the size of the cluster a given node belongs to is correlated with the degree of the node. Thus, $Q_a(s)$ must be evaluated as $Q_a(s) = \sum_{\vec{k}} P_a(\vec{k}) Q_a(s|\vec{k})$, where $Q_a(s|\vec{k})$ is the probability that a node from network a of degree $\vec{k}$ belongs to a cluster of size s. The latter function satisfies

$$\begin{aligned} Q_a(s|\vec{k}) = & \sum_{n_a} \binom{k_a}{n_a} p^{n_a} (1-p)^{k_a - n_a} \sum_{n_b} \binom{k_b}{n_b} p^{n_b} (1-p)^{k_b - n_b} \\ & \sum_{s_1 \cdots s_{n_a}} G_{aa}(s_1|\vec{k}) \cdots G_{aa}(s_{n_a}|\vec{k}) \sum_{s'_1 \cdots s'_{n_b}} G_{ab}(s'_1|\vec{k}) \cdots G_{ab}(s'_{n_b}|\vec{k}) \\ & \delta_{s, 1+s_1+\cdots+s_{n_a}+s'_1+\cdots+s'_{n_b}}, \end{aligned} \tag{6.2}$$

where $G_{aa}(s|\vec{k})$ ($G_{ab}(s|\vec{k})$) is the probability to reach s other nodes by following a neighbour in network a (b). The generating function of $Q_a(s|\vec{k})$ can be written as

$$\hat{Q}_a(z|\vec{k}) = \sum_{s=0}^{\infty} Q_a(s|\vec{k}) z^s = z(1 - p + p\hat{G}_{aa}(z|\vec{k}))^{k_a} (1 - p + p\hat{G}_{ab}(z|\vec{k}))^{k_b}. \tag{6.3}$$

Functions $G_{aa}(s|\vec{k})$, $G_{ab}(s|\vec{k})$, $G_{ba}(s|\vec{k})$, and $G_{bb}(s|\vec{k})$ follow similar recurrence equations. Thus, their generating functions satisfy

$$\hat{G}_{aa}(z|\vec{k}) = z \sum_{\vec{k}} P_{aa}(\vec{k}'|\vec{k}) (1 - p + p\hat{G}_{aa}(z|\vec{k}))^{k'_a - 1} (1 - p + p\hat{G}_{ab}(z|\vec{k}))^{k'_b} \tag{6.4}$$

$$\hat{G}_{ab}(z|\vec{k}) = z \sum_{\vec{k}} P_{ab}(\vec{k}'|\vec{k}) (1 - p + p\hat{G}_{ba}(z|\vec{k}))^{k'_a - 1} (1 - p + p\hat{G}_{bb}(z|\vec{k}))^{k'_b} \tag{6.5}$$

$$\hat{G}_{ba}(z|\vec{k}) = z \sum_{\vec{k}} P_{ba}(\vec{k}'|\vec{k}) (1 - p + p\hat{G}_{aa}(z|\vec{k}))^{k'_a} (1 - p + p\hat{G}_{ab}(z|\vec{k}))^{k'_b - 1} \tag{6.6}$$

$$\hat{G}_{bb}(z|\vec{k}) = z \sum_{\vec{k}} P_{bb}(\vec{k}'|\vec{k}) (1 - p + p\hat{G}_{ba}(z|\vec{k}))^{k'_a} (1 - p + p\hat{G}_{bb}(z|\vec{k}))^{k'_b - 1}, \tag{6.7}$$

where $P_{aa}(\vec{k}'|\vec{k})$ is the probability that a randomly chosen neighbour among all the a neighbours of a node that belongs to network a with degree $\vec{k}$ has degree $\vec{k}'$, and analogously for the rest of the transition probabilities.

For networks with no degree-degree correlations, these transition probabilities simplify as

$$\begin{aligned} P_{aa}(\vec{k}'|\vec{k}) &= \frac{k'_a P_a(\vec{k}')}{\bar{k}_{aa}} \quad P_{bb}(\vec{k}'|\vec{k}) = \frac{k'_b P_b(\vec{k}')}{\bar{k}_{bb}} \\ P_{ab}(\vec{k}'|\vec{k}) &= \frac{k'_a P_b(\vec{k}')}{\bar{k}_{ba}} \quad P_{ba}(\vec{k}'|\vec{k}) = \frac{k'_a P_a(\vec{k}')}{\bar{k}_{ab}}. \end{aligned} \tag{6.8}$$

This implies that functions $G_{aa}(z|\vec{k})$, $G_{ab}(z|\vec{k})$, $G_{ba}(z|\vec{k})$, and $G_{bb}(z|\vec{k})$ become independent of $\vec{k}$. We further assume that the number of neighbours from a and b of a given node are uncorrelated, that is

$$P_a(\vec{k}) = P_a(k_a)P_a(k_b) \quad P_b(\vec{k}) = P_b(k_a)P_b(k_b). \tag{6.9}$$

In the case of two coupled Erdős-Rényi random graphs, the degree distributions $P_a(k_a)$, $P_a(k_b)$, $P_b(k_a)$, and $P_b(k_b)$ are all Poisson distributions of parameter $\bar{k}_{aa}$, $\bar{k}_{ab}$, $\bar{k}_{ba}$, and $\bar{k}_{bb}$, respectively. In this case, it is easy to check that $\hat{Q}_a(z) = \hat{G}_{aa}(z)$, $\hat{Q}_b(z) = \hat{G}_{bb}(z)$, and

$$\hat{G}_{aa}(z) = ze^{-\bar{k}_{aa}p(1-\hat{G}_{aa}(z))}e^{-\bar{k}_{ab}p(1-\hat{G}_{ab}(z))} \tag{6.10}$$

$$\hat{G}_{ab}(z) = ze^{-\bar{k}_{ba}p(1-\hat{G}_{ba}(z))}e^{-\bar{k}_{bb}p(1-\hat{G}_{bb}(z))} \tag{6.11}$$

$$\hat{G}_{ba}(z) = ze^{-\bar{k}_{ab}p(1-\hat{G}_{ab}(z))}e^{-\bar{k}_{aa}p(1-\hat{G}_{aa}(z))} \tag{6.12}$$

$$\hat{G}_{bb}(z) = ze^{-\bar{k}_{bb}p(1-\hat{G}_{bb}(z))}e^{-\bar{k}_{ba}p(1-\hat{G}_{ba}(z))}. \tag{6.13}$$

Then, to calculate the fraction of nodes that belong to the giant component we use that $g_a = 1 - \hat{Q}_a(z=1) = 1 - \hat{G}_{aa}(z=1)$, $g_b = 1 - \hat{Q}_b(z=1) = 1 - \hat{G}_{bb}(z=1)$ and we also define $g_{ab} = 1 - \hat{G}_{ab}(z=1)$ and $g_{ba} = 1 - \hat{G}_{ba}(z=1)$. Then if we want network a to play the role of the core (c), so network b becomes the periphery (p), we have that $\bar{k}_{aa} = \bar{k}_c > \bar{k}_p = \bar{k}_{bb}$. In this new notation, we obtain the following system of transcendent equations:

$$\left.\begin{aligned} g_c(p) &= 1 - e^{-p\bar{k}_c g_c(p) - p\bar{k}_{cp} g_{cp}(p)} \\ g_{cp}(p) &= 1 - e^{-p\bar{k}_{pc} g_{pc}(p) - p\bar{k}_p g_p(p)} \\ g_{pc}(p) &= 1 - e^{-p\bar{k}_{cp} g_{cp}(p) - p\bar{k}_c g_c(p)} \\ g_p(p) &= 1 - e^{-p\bar{k}_p g_p(p) - p\bar{k}_{pc} g_{pc}(p)} \end{aligned}\right\}. \tag{6.14}$$

From here, it readily follows that g_c and g_p must be either both different from zero or equal to zero, implying that there is generally only one percolation transition, whereas at $p \approx \bar{k}_p^{-1}$, there is a crossover effect due to growth of the periphery.

This result is true if the coupling between the core and periphery is macroscopic, that is, the number of connections between the two structures is proportional to the size of the system such that $\bar{k}_{cp}$ and $\bar{k}_{pc}$ are constants in the thermodynamic limit. Instead, suppose that the number of connections among nodes in the core and periphery scales sub-linearly with the system size, *i. e.*, as N^{α} with $0 < \alpha < 1$. In this case, $\bar{k}_{cp}$ and $\bar{k}_{pc}$ are zero in the thermodynamic limit: thus, g_c and g_p become decoupled in Eq. (6.14) such that g_c can be different from zero while $g_p = 0$. However, when both the core and periphery have a giant connected component as isolated networks, the combined network forms a single connected component because there is an infinite number of connections between each part.

The effect of such structure on bond percolation is as follows. When the bond occupation probability is increased from $p = 0$, the first phase transition occurs at $p = \bar{k}_c^{-1}$, where the core percolates in a giant component $G_c \sim \mathcal{O}(N)$. In the range $\bar{k}_c^{-1} < p < \bar{k}_p^{-1}$, the periphery is composed of a large number of small disconnected components. The number of such components directly connected to G_c and, thus, the number of nodes in the periphery connected through G_c, scales as N^{α}; therefore, its fraction vanishes in the limit $N \gg 1$ and the relative size of the giant component of the combined system is just G_c/N. Once we reach $p = \bar{k}_p^{-1}$, a percolating cluster is formed in the periphery that becomes macroscopic as we increase p by an infinitesimal amount, *i. e.*, $G_p \sim \mathcal{O}(N)$. At this moment, and not before, the number of connections between G_c and G_p become $N^{\alpha-2}G_cG_p \sim \mathcal{O}(N^{\alpha})$ and, consequently, G_c and G_p are connected with probability 1. Thus, we have a double percolation transition defined by a regular transition at $p = \bar{k}_c^{-1}$ and the sudden emergence at $p = \bar{k}_p^{-1}$ of a macroscopic subgraph in the periphery with two types of connectivity; namely, each pair of nodes in this subgraph can be connected not only by a path going through the core but also by a path composed exclusively of nodes outside the core. In turn, this translates into a double discontinuity of the first (or higher) derivative of the order parameter g at $p = \bar{k}_c^{-1}$ and $p = \bar{k}_p^{-1}$, as clearly seen in Fig. 6.7 **b**.

Figures 6.7 **a, b** present the simulation results of the relative size of the giant component for $\alpha = 1$ and $\alpha = 0.5$, respectively. In the first case, we observe a crossover effect at approximately $p = \bar{k}_p^{-1}$ as also observed in [118], whereas in the second case, we observe a clear discontinuity in the derivative of $g(p)$ at exactly $p = \bar{k}_p^{-1}$, which is consistent with the analytical prediction in Eqs. (6.1) and (6.14) for $\bar{k}_{cp} = \bar{k}_{pc} = 0$. However, the strongest evidence for the presence of a genuine double phase transition is provided by analysis of the susceptibility. In the case of a crossover effect, fluctuations in the percolated phase behave as $\langle G^2 \rangle - \langle G \rangle^2 \sim \langle G \rangle$; consequently, the quantity χ should diverge at the critical point and become size-independent after this point has been surpassed. In contrast, if the second transition in the periphery is a real phase transition, this

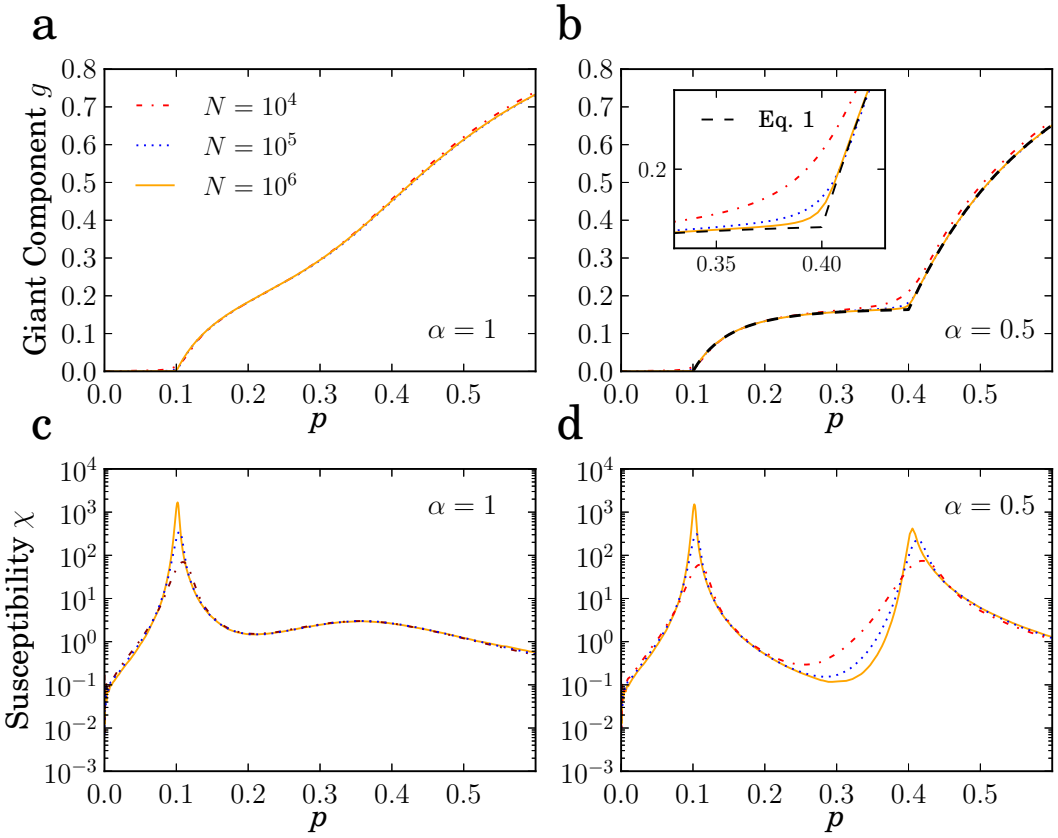


Figure 6.7: Bond percolation simulations for the core-periphery random graph model with $\alpha = 1$ (left column) and $\alpha = 0.5$ (right column). In both cases, the core has an average degree of $\bar{k}_c = 10$ and the periphery $\bar{k}_p = 2.5$. The core/periphery ratio is $r = 0.2$. The black dashed line in plot **b** is the numerical solution of Eqs. (6.1) and (6.14) with $\bar{k}_{cp} = 0$. The inset shows the approach to the theoretical prediction at the second transition point as the size of the system is increased.

quantity should diverge at both $p = \bar{k}_c^{-1}$ and $p = \bar{k}_p^{-1}$. This behaviour is clearly observed in Figs. 6.7 **c, d**.

In the case of clustered networks, it is difficult to clearly identify the core. Nevertheless using the giant $m1$-core as a rough approximation, we find that, in the case of $\bar{c}(k) = 0.25$, the average number of connections between a node not in the giant $m1$-core and nodes in the giant $m1$-core is approximately 0.02, indicating that the core and periphery are in fact very weakly coupled. In any case, the double divergence of χ shown in Fig. 6.3 **c**, just as in the core-periphery random graph model with $\alpha < 1$, is clear evidence for a genuine double phase transition.

6.8 Finite size scaling of the core-periphery random graph model

Let $(\beta_c, \gamma'_c, \nu_c)$ and $(\beta_p, \gamma'_p, \nu_p)$ be the critical exponents of the core and the periphery when they are isolated from each other. Close to the percolation transition of the core, the giant component is mainly composed of nodes in the core and, therefore, we expect the first transition to have the critical properties of regular percolation in the core subgraph, in particular, the susceptibility near the first peak diverges with the exponent γ'_c/ν_c. Close to the second transition point, the giant component is the sum of the giant component in the core, G_c, plus the percolating cluster in the periphery, G_p. Since G_c and G_p are statistically independent, the susceptibility in this region can be evaluated as

$$\chi = \frac{\langle G_c\rangle\chi_c + \langle G_p\rangle\chi_p}{\langle G_c\rangle + \langle G_p\rangle} \approx \chi_c + \frac{\langle G_p\rangle}{\langle G_c\rangle}\chi_p. \tag{6.15}$$

However, if the second transition point is well separated from the first one, close to this second transition $\chi_c \sim$ cte, $\langle G_c\rangle \sim N$, and $\langle G_p\rangle \sim N^{1-\beta_p/\nu_p}$. Then, we expect that near the second transition the susceptibility behaves as $\chi \sim N^{(\gamma'_p-\beta_p)/\nu_p}$. The critical exponents in the case of Erdős-Rényi random graphs are the mean-field ones, that is, $\beta = \gamma = 1$ and $\nu = 3$ [165] (See appendix A.2 for more detailed explanation of the mean field values) . Therefore, in our simulations, we expect the first peak to diverge as $N^{2/3}$, the second peak as $N^{1/3}$ and the approach of the position of the peaks to their respective critical points as $p_{max} \sim p_c + AN^{-1/3}$. This is confirmed in Fig. 6.9.

6.9 Double percolation in real networks

The multiple percolation phase transition phenomenon that we just reported, its not exclusive of our synthetic networks. We performed an extensive anal-

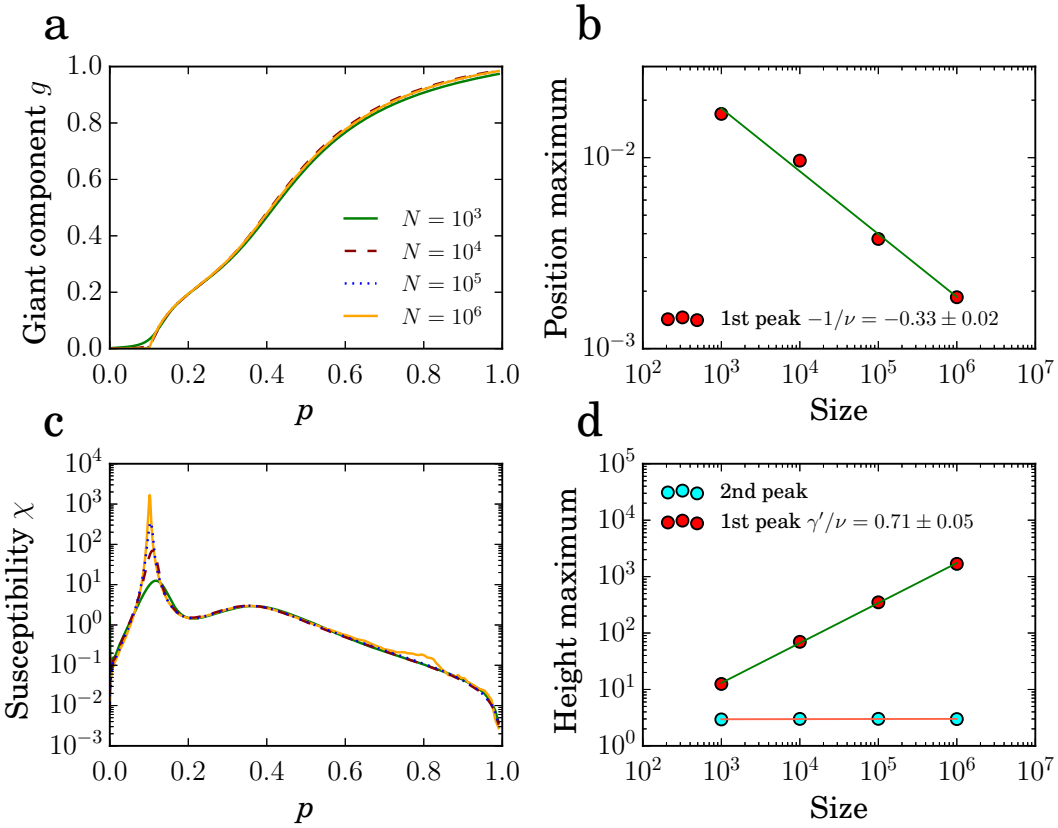


Figure 6.8: Bond percolation simulations for the core-periphery random graph model for $\alpha = 1$ for different sizes. In both cases the core has an average degree $\bar{k}_c = 10$ and the periphery $\bar{k}_p = 2.5$. The ratio core/periphery is $r = 0.2$. **a:** Relative size of the largest connected component as a function of the bond occupation probability p. **c:** Susceptibility χ as a function of bond occupation probability p. **b** and **d:** Position p_{max} and height χ_{max} of the two peaks of χ as function of network size N. The straight lines are power-law fits and the measured values of the critical exponents are shown.

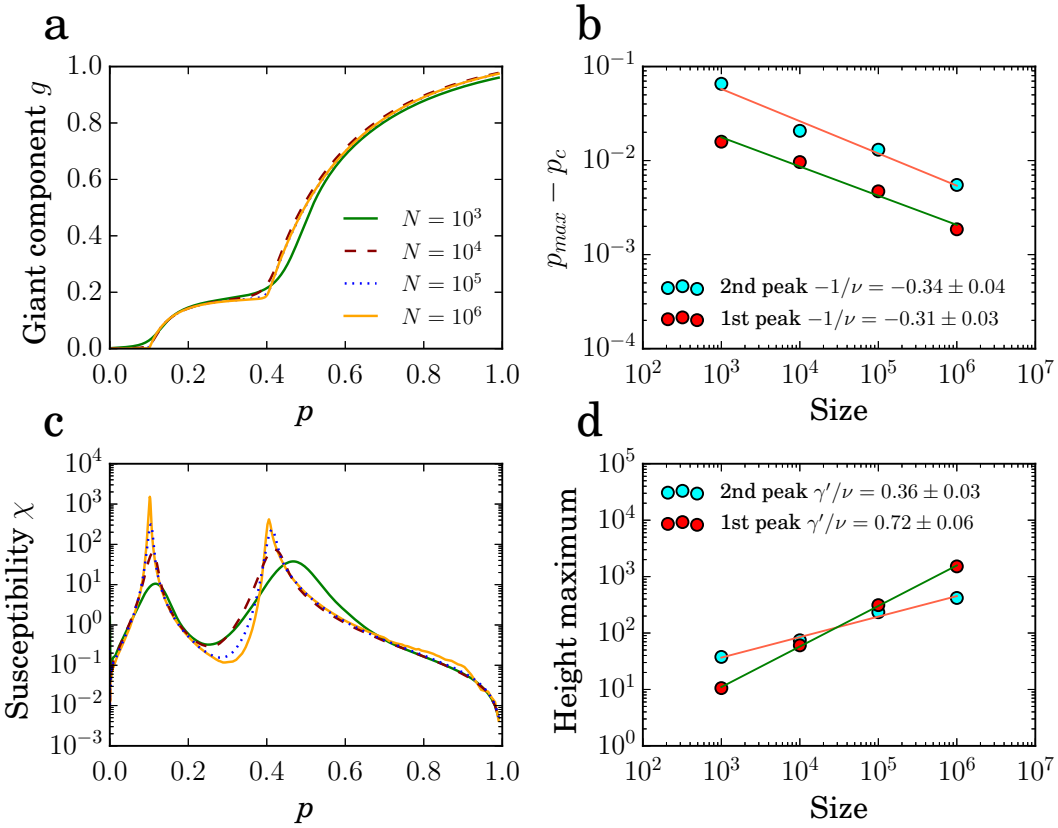


Figure 6.9: Bond percolation simulations for the core-periphery random graph model for $\alpha = 0.5$ for different sizes. In both cases the core has an average degree $\bar{k}_c = 10$ and the periphery $\bar{k}_p = 2.5$. The ratio core/periphery is $r = 0.2$. **a:** Relative size of the largest connected component as a function of the bond occupation probability p. **c:** Susceptibility χ as a function of bond occupation probability p. **b** and **d:** Position p_{max} and height χ_{max} of the two peaks of χ as function of network size N. The straight lines are power-law fits and the measured values of the critical exponents are shown.

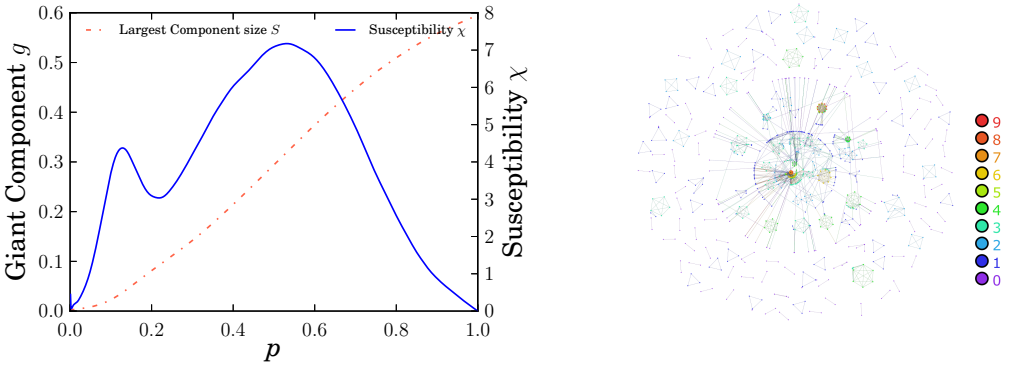


Figure 6.10: Left: Bond percolation simulations for the human disease network network. The relative size of the largest connected component g and its susceptibility χ as a function of the bond occupation probability p. Right:m-core decomposition

ysis of the bond percolation properties of many real networks and we found many examples in which a multiple percolation transition may occur. For example, figure 6.10 shows the susceptibility of the giant component as a function of the bond occupation probability of the disease network and its m-core structure visualization (See appendix A.1.6, A.1.7, A.1.5 for a description of the network as well as other real examples). As we can observe, the susceptibility shows two well defined peaks suggesting the presence of a multiple percolation phase transition. The m-core visualization of the disease network shows a clear core-periphery structure as observed in our synthetic networks.

However, giving the finite size of real networks and that we do not have access to the ensemble they belong, it is not possible to perform a finite size scaling. Therefore, any of the existing empirical methods allows to differentiate between a real phase transition from a smeared one so we can not know which of the peaks is a true percolation transition. Besides we can not guarantee that there are other percolation transitions that have no peak in the susceptibility due to finite size effects. This fact makes any of the empirical methods a good indicator of the position of the different percolation threshold of a network. In the next chapter we are going to address this problem by adapting the state-of-the-art theories to the new multiple percolation phenomenon.

6.10 Discussion

As we have demonstrated, clustering has a non-trivial effect on the properties of complex networks. This effect depends on three main factors: the heterogeneity of the degree distribution, the degree-degree correlations, and the shape of the clustering spectrum $\bar{c}(k)$. If we avoid degree-degree correlations, the combination of strong clustering and heterogeneity induces the emergence of a small but macroscopic core surrounded by a large periphery. This organization redefines the percolation phase space of complex networks by inducing a new percolated phase in which the core of the network is percolated but the periphery is not. In this situation, increasing clustering makes the core larger and more entangled, thereby decreasing the percolation threshold of the first transition, as suggested in [84, 132, 154]. However, in the remaining part of the network (the periphery) clustering generates small clique-like structures that are sparsely interconnected (see Fig. 6.4 c). Thus, the periphery becomes more fragile, and the percolation threshold of the second phase transition increases, in agreement with [86, 104, 122, 134]. For weakly heterogeneous networks, the size of the core is not macroscopic; thus, clustering only makes these networks more susceptible to the removal of links. This fact reconciles the two dominant interpretations of the effect of clustering on the percolation properties of complex networks. Interestingly, this behaviour is also observed in a large sample of real complex networks, which provides evidence of the generality of this phenomenon.

We have shown that, in contrast to previous theory, it is possible to have two or more consecutive continuous phase transitions associated with the same symmetry breaking. Our work opens new lines of research concerning the effect of this core-periphery architecture on dynamical processes that occur in networks. In the case of epidemic spreading, for instance, the core could act as a reservoir of infectious agents that would be latently active in the core while the remainder of the network is uninfected.

CHAPTER 7

Local percolation thresholds

7.1 Network of networks

The common conception in complex networks is that there is a single percolation threshold, p_c, that depends on the whole structure of the network concerned. It is therefore considered extremely important to have an accurate value of p_c. In the case of infrastructure or technology networks, this critical point defines the threshold between global functioning and total collapse. In epidemiology, the epidemic threshold determines whether a disease will die out or reach an endemic steady state. Hence, much effort has been devoted to develop theoretical and numerical methods to find an accurate measure of the percolation threshold.

However, in the previous chapter we demonstrated that, in contrast to this common belief, networks can undergo more than one percolation phase transition. We showed that this phenomenon occurs when different weakly connected parts of the network percolate independently at their own critical points. This radically changes the percolation phase space of complex networks but, what is more important, it redefines how to tackle the percolation problem. In this new scenario, due to random failure of connections, we can have a finite fraction of the network that is completely fragmented while the rest of the network remains perfectly functional. In epidemiology, this implies that a disease can become pandemic in some areas of the network, while other parts remain healthy. So, it is no longer in our best interests to try to find the unique percolation threshold of a network, but rather we should look for a set of critical points at which different parts of the network percolate.

The possibility that a network has different parts, or modules, that are independent in terms of network processes, suggests that we could consider each module as a network itself and the whole system as a network of networks (NON) [19, 35, 79]. In such a situation, each module can have different structural organization, while the coupling between different modules/networks determines the global properties of the whole system. Obviously, in the case of real networks, it is extremely difficult to determine whether a given network can be cast within this framework of a NON or whether it is better represented by a more traditional description in terms of communities. In this chapter, we study

the percolation properties of real networks and based on that study we present a method that can establish whether a specific network under consideration is a NON.

7.2 Measure multiple percolation thresholds

The existence of multiple percolation phase transitions has important practical implications for the numerical methods to measure the percolation threshold. In real networks, for a given value of the bond occupation probability p, the percolation process is repeated a large number of times and for each such realization, the size of the largest (or giant) component, G, is measured [137]. The average of this giant component over different realizations, $\langle G \rangle$, and its fluctuations (the susceptibility), $\chi(p)$, given by

$$\chi(p) = \frac{\langle G^2 \rangle - \langle G \rangle^2}{\langle G \rangle}, \tag{7.1}$$

encode the critical properties of the percolation transition [165]. In finite systems, a peak in $\chi(p)$ indicates the presence of a continuous phase transition and its position provides what is believed to be the best estimate of the critical point, p_c [148].

However, as we have seen, in many real networks, the susceptibility shows two or even more peaks; and in such cases, the information encoded in $\chi(p)$ related to the critical properties of the system is unclear, to say the least. Indeed, multiple peaks in the susceptibility are present in NONs when the number of links between the different networks scales sub-linearly with the system size. In such cases, these multiple peaks indicate the existence of multiple percolation transitions corresponding to the individual percolation transition of each network [49]. However, one can also find multiple peaks when the coupling between the different networks is linear. In this latter case, the system is better described as a single network, possibly with a marked community structure, and of all the peaks in $\chi(p)$, only one is actually related to the critical percolation transition.

Ideally, to distinguish the two cases, one should perform finite-size scaling analysis, where peaks associated with real phase transitions should diverge as the system size increases, or remain constant otherwise. Unfortunately, this approach is infeasible in real networks, for which a single fixed-size network is available. Thus, given a real network with multiple peaks in the susceptibility, none of the existing numerical methods allows us to discern which of them corresponds to a true percolation transition. Even worse, as we show below, even when the susceptibility exhibits just a single peak, it is not possible to guarantee

that there are no other phase transitions that may be hidden, due to finite size effects. Therefore, the possibility that a real network may undergo more than one percolation transition makes the traditional numerical methods for measuring the percolation threshold obsolete for real networks. Moreover, the lack of genuinely numerical methods that can identify all the percolation thresholds leaves us with no mechanism to check the reliability of theoretical values of the percolation threshold.

Much effort has been devoted to obtaining an accurate analytic expression for the bond percolation threshold [33, 37, 46, 99]. Systematic analysis using real networks to quantify the reliability of these analytic expressions has shown that the inverse of the largest eigenvalue of the non-backtracking matrix (NBTM) represents the best measure currently available [148]. However, that measure was found to be less accurate for large values of the true percolation threshold. The reason for this poor performance was argued to reside in the localization of the principal eigenvector; the percolation transition in reality involves only a finite fraction of the nodes in the network and thus does not correspond to the true percolation transition. An analogous explanation was found for spectral estimators of epidemic thresholds in real networks [88]. However, those authors were not aware of the multiple percolation transition phenomenon and the limitations of numerical methods to measure the percolation threshold associated with it. They assumed that there was only one critical point, p_c, and that its best estimate was the value of p, where the susceptibility reaches its maximum, irrespective of whether there were other local maxima.

Here, we adapt the most recent percolation theory [99] to the multiple percolation framework. We show that the localization of the principal eigenvector of the NBTM can indicate the existence of multiple percolation transitions. In this situation, all the critical points correspond to the inverse of one eigenvalue, and the leading eigenvalue only gives us the position of the first phase transition. This critical point may not coincide with the position of the global maximum of the susceptibility because the later may correspond to another transition. Furthermore, we use the theory developed in [99] to classify all nodes by a new measure that can be interpreted as the percolation threshold of each node p_{ci}. According to this classification, we can now identify the critical points of all the phase transitions that a network may have, and the nodes involved in each one of them. This allows us to reveal the internal organization of a network into a NON. We consider that our work offers new insight into the concept of percolation threshold, which was initially conceived as a macroscopic magnitude and now becomes a local property, and how this affects the numerical methods to identify it and the accuracy of the theoretical values.

7.3 The node percolation threshold

The theory developed in [99] uses a message passing technique to calculate, from a given adjacency matrix, the probability $H_{i\leftarrow j}$ that node i does not belongs to the giant component (GC) through the edge that connects it to node j. Using the tree-like assumption and the generating function approach one can express the probability $H_{i\leftarrow j}$ as:

$$H_{i\leftarrow j} = 1 - p + p \prod_{k\in\mathcal{N}_j\setminus i} H_{j\leftarrow k}, \tag{7.2}$$

where $\mathcal{N}_j \setminus i$ defines the set of nodes in the direct neighbourhood of node j excluding node i, encoded in the adjacency matrix. The exclusion of node i is to avoid backtracking messages. Then, the probability g_i that node i belongs to the GC is then:

$$g_i = 1 - \prod_{j\in\mathcal{N}_i} H_{i\leftarrow j}, \tag{7.3}$$

so the fraction of nodes that belong to the giant component, g, can be expressed as a simple average of g_i over all the nodes.

To find the percolation threshold, the stability of the trivial solution $H_{i\leftarrow j} = 1$ is studied. By introducing a small perturbation, $\epsilon_{i\leftarrow j}$, and linearising Eq. 7.2, the authors obtain the equation: $\boldsymbol{\epsilon} = p\boldsymbol{B}\boldsymbol{\epsilon}$, where $\boldsymbol{B}$ is the non-backtracking matrix (NBTM). It is a non-symmetric matrix with rows and columns indexed by directed edges, $i \leftarrow j$, and elements, $B_{i\leftarrow j,k\leftarrow l} = \delta_{jk}(1-\delta_{il})$. The fixed point is stable under iteration if and only if p times the leading eigenvalue of $\boldsymbol{B}$ is less than unity. Hence, they conclude that the critical percolation probability, p_c, is equal to the inverse of the leading eigenvalue of $\boldsymbol{B}$. Therefore, for a globally connected network, all the nodes have a non-zero probability of belonging to the GC at $p > p_c$, so the network percolates at p_c.

However, this is not the case when we have a bloc diagonal NBTM; which corresponds to a system composed of two disconnected networks. Edges from one network will have a zero component in the eigenvectors localized in the other network. Therefore, such a system undergoes two percolation phase transitions and their critical points will be at the inverse of the two leading eigenvalues of the principal eigenvectors localized in each network. A slightly different problem arises when the two networks are connected. In this case, depending on the number of interlinks, they can either percolate at once or percolate independently. As we show in chapter 6, a double percolation phase transition is still possible when the inter average degree vanishes in the thermodynamic limit [49]. A network with such a configuration will have an almost bloc diagonal NBTM with eigenvectors localized in each network. Edges from one network will have a very small component in the eigenvectors localized in the other

network. In fact, these components will vanish in the thermodynamic limit as soon as their value is proportional to the inter average degree. Therefore, a finite network generated from this ensemble will have, mathematically speaking, only one percolation threshold. However, in the thermodynamic limit, so physically speaking, this ensemble of networks has two phase transitions.

In conclusion, if the principal eigenvector of a real network is localized, this means that the network may have multiple percolation phase transitions. Thus, there is not a single true percolation threshold, but a set of them. In this situation, each critical point corresponds to an eigenvalue of the NBTM. The inverse of the leading eigenvalue only gives the position of the first percolation transition. In contrast, the position of the maximum of the susceptibility will give us any one of the critical points, which may or may not coincide with the smallest one. However, not all the eigenvectors are associated with a phase transition. Therefore, simple spectral analysis of the NBTM is not an effective method to identify how many transitions occur in a real network or to measure their critical points.

Finding a method that gives the different critical points of multiple phase transitions and the nodes involved in each one of them is analogous to finding the percolation threshold of each node, p_{ci}. This is possible if p_c is not a macroscopic property of the network but as a local one. If a network has only one transition, we expect all the nodes to have exactly the same p_{ci}. If the nodes have different percolation thresholds, this means that parts of the network percolate at different times.

According to Eq. 7.3, beyond the inverse of the leading eigenvalue, all the nodes will have a non-zero probability of belonging to the giant component. Therefore, given a finite system, we cannot use Eq. 7.3 to find the percolation thresholds of the nodes. Instead, we use the average size of finite clusters that each node belongs to. This magnitude is calculated in [99] as:

$$\langle S \rangle_i = 1 + \sum_{j \in \mathcal{N}_i} \frac{H'_{i \leftarrow j}}{H_{i \leftarrow j}}. \tag{7.4}$$

where $H'_{i \leftarrow j}$ can be expressed as:

$$H'_{i \leftarrow j} = p \left[1 + \sum_{k \in \mathcal{N}_j \setminus i} \frac{H'_{j \leftarrow k}}{H_{j \leftarrow k}} \right] \prod_{k \in \mathcal{N}_j \setminus i} H_{j \leftarrow k}. \tag{7.5}$$

In classical percolation, at the critical point, the average finite cluster size of the whole system reaches its maximum. Beyond this point, adding more edges does not increase the size of the finite clusters, but merges the larger clusters with the percolating (infinite) one, the GC. Therefore, the position of the maximum

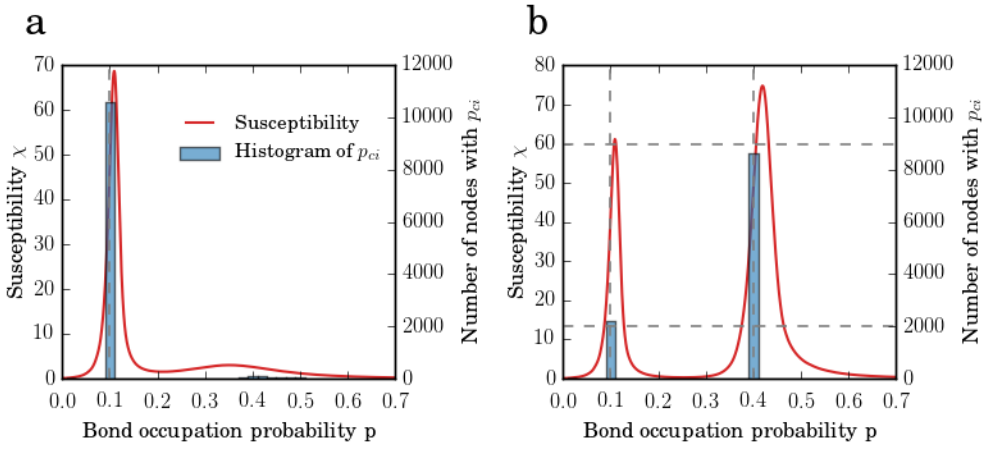


Figure 7.1: Bond percolation simulations for two interconnected networks with different sizes ($N_A = 2000$ and $N_B = 10000$) and inner average degree ($\bar{k}_A = 10$ and $\bar{k}_B = 2.5$). Left axis: susceptibility of the size of the giant component χ. Right axis: a histogram of the position of the maximum of the average size of finite clusters of all the nodes. **a:** Number of interlinks equal to the total number of nodes 12000 so there is one phase transition. **b:** Number of interlinks equal to the square root of the number of nodes 110. In this case we have two phase transitions.

of the average finite cluster size is used as an indicator of the percolation threshold [165].

By analogy, this should hold for each node. The position of the maximum of the average size of the finite clusters that a node belongs to indicates the point at which the node starts to belong to the GC, which corresponds to its percolation threshold.

To check the reliability of our measure, we use as a benchmark the simple model introduced in chapter 6. We create two Erdős-Rényi graphs, A and B, with different numbers of nodes ($N_A = 2000$ and $N_B = 10000$) and inner average degree ($\langle k \rangle_A = 10$ and $\langle k \rangle_B = 2.5$). Then we consider two scenarios. In the first, the number of interlinks is equal to the total number of nodes (12000), so there are enough interlinks to have only one phase transition and there will be only one critical point at $p_c = \frac{1}{\langle k \rangle_A} = 0.1$. In the second scenario, the number of interlinks is equal to the square root of the number of nodes (110), so the inter average degree of the nodes vanishes in the thermodynamic limit. Thus, networks A and B will percolate independently from each other at the critical points $p_{cA} = 0.1$ and $p_{cB} = 0.4$.

For both cases, we calculate the percolation threshold of each node, p_{ci}, as the position of p at which their average cluster size shows its maximum. Fig-

ure 7.1 shows, for both scenarios, the histogram of the percolation threshold of each node compared with the susceptibility χ of the whole system. In figure 7.1 **a**, where there is only one phase transition at $p_c = 0.1$, 95% of the nodes have exactly the same percolation threshold: $p_{ci} = p_{cA} = 0.1$. In the second scenario, (Fig. 7.1 **b**), there are 2204 nodes that percolates at the first critical point, $p_{cA} = 0.1$. The rest of the nodes percolate exactly at $p_{cB} = 0.4$, suggesting the presence of a multiple percolation transition. In this latter situation, the inverse of the leading eigenvalue of the NBTM is 0.1, although the global maximum of the susceptibility is 0.419. This does not imply that the theory is inaccurate, but that it only gives us the position of the first transition.

If we now look at the performance of our method in the classification of the nodes, we observe that they do not coincide perfectly with the original network partition. For example, in the first scenario, where all nodes should have the same percolation threshold, there are few nodes of network B with a percolation threshold equal to p_{cB}. In the system with two phase transitions, there are 204 nodes of network B that percolates at the first critical point, p_{cA}. This discrepancy comes from nodes on the border between the two networks which our method does not classify properly. However, if the connection between the networks is weak enough for them to percolate independently, the fraction of nodes in network B that belong to the border region will vanish in the thermodynamic limit; so the error in node classification is just a finite size effect. Nevertheless, our method can tell whether a given network it is better represented by a more traditional description in terms of communities or whether it is a NON. In the later case, despite some finite size effects, our method is able to determine which network of the NON each nodes belongs to.

7.4 Real networks

Once we have shown that our method works for the benchmark, we now proceed to analyse some real networks. The first network that we analyse is the Gnutella peer-to-peer file sharing network [110] (see appendix A.1 for a detailed description of the real networks used in this section). In Figure 7.2 **a**, we can see that the susceptibility of the giant component shows two well-defined peaks at $p_1 = 0.045$ and $p_2 = 0.104$. However, the histogram of the percolation thresholds of the nodes shows that 99.4% of the nodes have the same percolation threshold: $p_c = 0.038$. Hence, our method reveals that there is only one phase transition and that only the first peak of the susceptibility is a true critical point. Therefore, the Gnutella network should be considered as a network with a marked community structure, rather than a NON.

Furthermore, being able to identify which of the peaks in the susceptibility

correspond to true critical points allows us to analyse the accuracy of the theoretical values of the percolation threshold. In this case, the inverse of the leading eigenvalue of the NBTM is $p_c = 0.038$; while the real bond percolation threshold is $p_c = 0.045$.

The second network we analyse is the human disease network in which nodes represent disorders and two disorders are connected if they share at least one gene in which mutations are associated with both disorders [87]. Figure 7.2 **b** shows that the susceptibility of the size of the GC of the disease network has two well-defined peaks. If we look at the histogram of the percolation thresholds of the nodes, we observe that nodes have different percolation thresholds. 196 out of the 516 nodes in the network ($\sim 38\%$) have the percolation threshold at $p = 0.088$, which corresponds to the inverse of the leading eigenvalue. The rest of the nodes group into very small modules with percolation thresholds that range from 0.1 to 0.4. This implies that the disease network is undergoing not one, but many phase transitions; so it is made of weakly connected modules. Therefore, the disease network would best be treated as a NON, instead of as a network with a community structure.

To confirm the modular structure of the network, Figure 7.3 **a** is a plot of the adjacency matrix with the nodes labelled according to the percolation threshold of each one, in ascending order. Fig 7.3 **a** shows that the disease network is indeed made up of a large module that is weakly connected to many smaller modules. This adjacency matrix is bloc diagonal and therefore so is the NBTM, which implies that its principal eigenvector is localized in the largest module.

We then isolate the largest module, with the smallest percolation threshold, $p_c = 0.088$, to perform independent bond percolation simulations and we compare the results with the original network. Figure 7.3 **b** shows that the largest module is responsible for the first peak in the susceptibility of the disease network at $p_c = 0.18$. Interestingly, 90% of the of all the disorders related to cancer populate almost half of this module.

Before considering the multiple percolation transition phenomenon, the best estimate of the percolation threshold of the disease network was considered to be: $p_c = 0.529$, which corresponds to the global maximum of the susceptibility. The theoretical value given by [99] is $p_c = 0.088$, so it was considered that the theory performed poorly in this case. However, our analysis reveals that the maximum of the susceptibility does not correspond to the transition related to the principal eigenvector of the NBTM but to the percolation of all the small modules connected to it. This fact implies that the theory did not perform poorly but that we were only considering the principal eigenvector, that is localized in the large module. Therefore, the theoretical value of $p_c = 0.088$ given by the inverse of the principal eigenvector of the NBTM corresponds to the first peak of the susceptibility at $p_c = 0.18$.

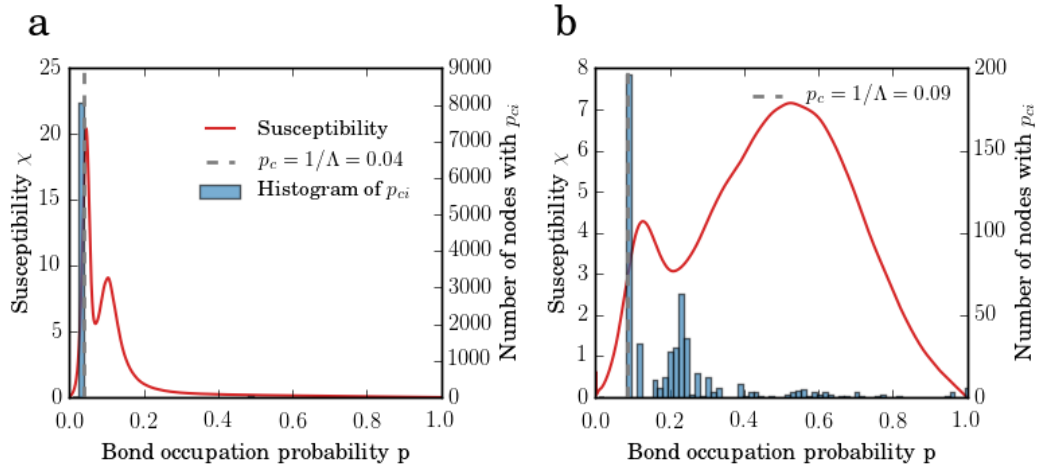


Figure 7.2: Bond percolation simulations for the Gnutella file sharing network (**a**) and the human disease network (**b**). Left axis: the susceptibility of the size of the giant component, χ. Right axis: a histogram of the position of the maximum of the average size of finite clusters of all the nodes.

The third network we analyse is a science collaboration network in which network scientists are connected by co-authorship of papers [133]. Figure 7.4 **a** shows that the susceptibility of the giant component of the network has one well-defined peak at $p = 0.406$. However, the histogram of the percolation thresholds of the nodes reveals that there are multiple transitions. Among the different modules, there is a large module (47%) with a critical point that coincides with the inverse of the leading eigenvector ($p = 0.11$). However, in real simulations, the different transitions have such similar percolation thresholds that they all merge into one peak of the susceptibility.

If we plot the adjacency matrix with the node labels ordered according to the percolation threshold of each node, we can see that the adjacency matrix is bloc diagonal, in agreement with the necessary condition for multiple percolation transitions (Fig. 7.4 **b)**. Hence, the network science collaboration network should be cast within the NON framework. In this situation, it is known that the appearance of coexistent percolating clusters can cause significant error in the message passing percolating theory [70]. This is an example in which our method has revealed that a real network has a much more complex percolation process than expected using traditional percolation analysis.

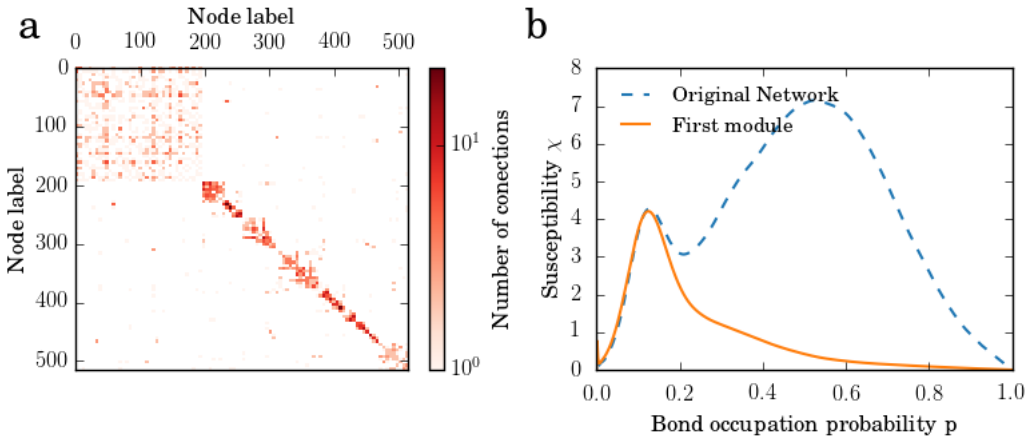


Figure 7.3: **a)** Adjacency matrix of the human disease network, with the node labels ordered according to the percolation threshold of each node, in ascending order. **b)** Independent bond percolation simulations of the human disease network and the module with nodes with the smallest percolation threshold $p_c = 0.088$. The susceptibility of the size of the giant component, χ, as a function of the bond occupation probability.

7.5 Discussion

Networks with more than one percolation transition do not have a true percolation threshold, but a set of critical points. The multiple phase transitions correspond to weakly connected modules that percolate independently of each other. In this scenario, different parts of the network can be completely fragmented while the rest of the network remains functional. This type of percolation process is much more rich and complex than the previous conception of a single phase transition, and this has important implications for infrastructure networks and epidemiology.

In this situation, none of the existing numerical methods is capable of properly identifying the set of critical points that a network has. The susceptibility of the giant component, χ, can have more than one peak and none of the existing numerical methods allows us to discern which of them corresponds to a true percolation transition. Moreover, theoretical effort still focuses on finding a unique true percolation threshold and has not yet adapted to this new scenario.

In this work, we calculate the percolation threshold of each node by looking at the maximum of the average cluster size that every node belongs to, using message passing theory [99]. By classifying all the nodes according to this new

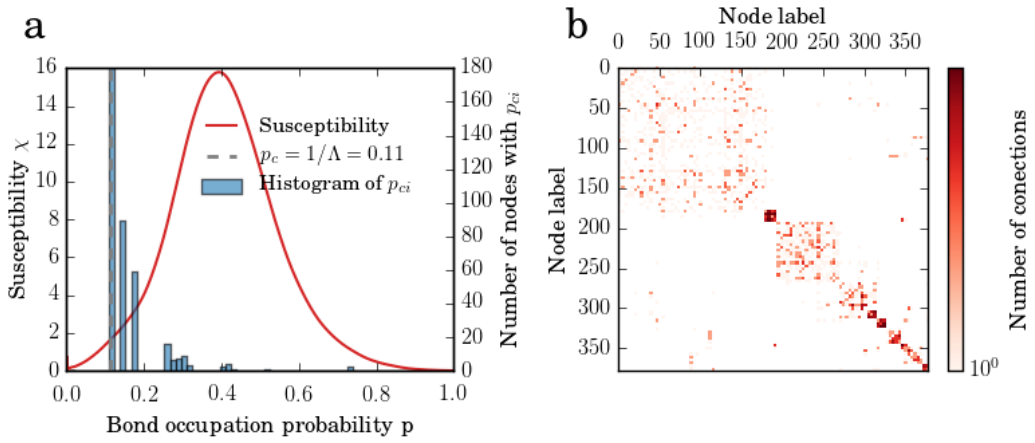


Figure 7.4: **a)** Bond percolation simulations of the network collaboration network. Left axis: the susceptibility of the size of the giant component, χ. Right axis: a histogram of the position of the maximum of the average size of finite clusters of all the nodes. **b)** Adjacency matrix of the network with the node labels ordered according to the percolation threshold of each node.

measure we have been able to identify which of the peaks of the susceptibility are smeared transitions and which are real critical points. We analyse three real cases in which our method reveals a modular structure, with different effects on the bond percolation properties in each case. In the disease network case, we were able to identify which part of the network is responsible for each peak observed in the susceptibility of the giant component. In contrast, in the Gnutella file sharing network, we were able to detect that the second peak in the susceptibility was a smeared transition and that there was only one true percolation critical point. In the science collaboration network, we found multiple percolation transitions, although the susceptibility showed only one peak.

Moreover, our measure also reveals the modular structures of networks that undergo multiple percolation phase transitions. The bloc diagonal adjacency matrices imply that the principal eigenvector of the NBTM is localized. In such a situation, the critical point of each module corresponds to an eigenvalue. This, together with [109], renews interest in spectral analysis of the non-backtracking matrix.

Our work contributes to establishing whether a given network should be cast within the NON framework or whether instead it is best represented by a more traditional description in terms of communities. Furthermore, we consider that these results have important applications in bond percolation theory, which

nowadays is still focused on finding the unique and true percolation threshold.

CHAPTER 8

Summary and conclusions

The study of a system from a network perspective focuses on the impact that connectivity between the elements has on the function of the system. The observation and measurement of parameters of real-world networks reveals that these systems have highly complex structures that differ from those of lattices and random graphs, and which have striking effects on their behaviour. Moreover, some common topological properties shared by networks with completely different natures have been found. This suggests the existence of common fundamental principles that determine the structure and evolution of networks.

One of the most common features of real networks is the high presence of triangles or strong clustering. However, in contrast to other topological properties of real networks, little was known about the emergence of clustering and its effect on network structure and function. The reason for this was twofold. First of all, the mere presence of triangles in networks contradicts assumptions that are used almost across the board in mathematical tools that are applied in network theory, and therefore it hinders any analytical treatment. Second, there was a lack of appropriate clustered network models that allow empirical study. Therefore, clustering was the main factor that thwarted the possibility of applying network theories to real situations and became one of the most important challenges facing network science.

In this thesis I study the role played by clustering in the structure and function of complex networks. To that end, I first analyse the clustering generated by random network models. Then, I study the organization of triangles within real networks. Finally, I focus on the effect of clustering on the classical bond percolation problem. My choice was based on the direct relation that this simple process has with robustness and the epidemics dynamics of networks. In this way a primary question arose: How does clustering affect the position of the bond percolation threshold?

Percolation in clustered networks had been widely studied before. However, previous reports did not agree on the results concerning the position of the percolation threshold. Although all the work was correct, much was only valid for a specific network structure; which I show here not to be similar to the structure of real networks. Therefore, I develop an appropriate clustered network model that reproduces the global organization of triangles in real networks better than previous clustered network models. Finally, I use my new model to study how

clustering actually affects the position of the bond percolation threshold of networks.

Results

In chapter 2 I introduce the exponential random graph models that generate maximally random networks with a given set of constrains, which are fixed on average. From this collection of ensembles, I focus on maximally random graphs with an expected degree sequence. This model is the canonical ensemble of one of the most popular models in network theory: the configuration model. Fixing the expected degree of all the nodes instead of the actual degree allows for an analytical treatment that is not possible in the micro-canonical ensemble.

Within this framework, in chapter 3, I present an analytical study of the clustering generated by random scale-free networks, and amend previous incorrect results for highly heterogeneous networks. I found the correct scaling behaviour of the clustering coefficient of the ensemble of scale-free random graphs with exponent $2 < \gamma < 3$ given by Eq. 3.11. Interestingly, for values of the exponent $\gamma \approx 2$, clustering remains nearly constant up to extremely large network sizes, but it is not self-averaging. This implies that highly heterogeneous networks can have a moderate level of clustering only due to topological constrains given by the degree distribution. This contradicts the common believe that random networks can always be approximated by a tree network and thereby violates a very common assumption in network theory.

However, the clustering generated in random networks is still not comparable to that observed in many empirical networks. Therefore, there was a need for clustered network models that could be used to study the effect of clustering on other topological properties of networks and processes on them. Along these lines, in section 2.5, I develop a model that generates clustered network models from an exponential random ensemble via a biased rewiring procedure.

My clustered network model has two important features that make it more convenient than other models. First, it can give different levels of clustering while fixing both the degree distribution and degree correlations. This is an important issue in order to disentangle the effect of clustering from the other two topological properties. Second, my model is an exponential graph and therefore maximizes the entropy of the network, making only minimum assumptions other than those imposed by the constraints. Therefore, in my model, in contrast to other clustered network models, the distribution of triangles in the network is as random as possible.

Furthermore, I released the code RandNetGen [48] that generates networks using my clustered network ensemble. The program goes beyond my model and

can generate any network from an exponential random graph ensemble using the biased rewiring method and taking many different topological properties as constraints. The program is user friendly and is published in the collaborative open-source platform Github.

Then, in chapter 4, I consider the distribution of triangles in real networks. To that end, I study the m-core structure, which is much deeper if triangles are distributed at random than it is if they are in a modular structure, in which strong correlation among multiplicity of edges is present. Moreover, I develop a visualization tool, LaNet-Vi 3.0 [16], that provides visual representations of the m-core structure of a network. My results show that the global organization of clustering in real networks is much better reproduced by my maximally random model than by previous clustered network models, in which triangles are ordered in a very specific way. Therefore, the good performance of my model defines the proper framework for studying the effect of clustering on bond percolation properties.

Afterwards, I use my model to study how clustering affects the percolation properties of networks. To do this, I compared the bond percolation properties of networks with the same degree sequence and degree correlations but different levels of clustering.

My results show that the effect of clustering depends strongly on network heterogeneity. For weakly heterogeneous networks, clustering increases the percolation threshold, thereby making the networks more fragile. However, for more heterogeneous networks ($\gamma \leq 3.5$) an increase in clustering can induce the emergence of a core–periphery structure. This organization redefines the percolation phase space of complex networks by inducing a new phenomenon, namely a double percolation phase transition, in which the core and periphery percolate independently.

In this situation, increasing clustering decreases the percolation threshold of the core and increases the percolation threshold of the periphery. For weakly heterogeneous networks, the size of the core is not macroscopic; thus, clustering only makes these networks more susceptible to the removal of links. This reconciles the two dominant interpretations of the effect of clustering on the percolation properties of complex networks.

Furthermore, this multiple percolation phase transition phenomenon that I reveal completely redefines our previous understanding of percolation process on complex networks. In previous theory, it was not possible to have two or more consecutive continuous phase transitions associated with the same symmetry breaking. Nevertheless, in section 6.7, I analytically prove that such anomalous transitions are indeed possible. I show that two weakly connected macroscopic modules of a network can percolate independently as long as their inter average degree scales sub-linearly with the system size. Interestingly, this behaviour is

also observed in many real complex networks; evidence of the generality of this phenomenon.

Therefore, due to the considerable heterogeneity of real networks, we no longer have a true percolation threshold but a set of critical points at which different parts of the network percolate. This changes the way we have to tackle the percolation problem and requires a reinterpretation of all the theoretical effort to find a single percolation threshold.

The possibility that a network has different parts, or modules, that are independent in terms of network processes, suggests that we should consider each module as a network itself and the whole system as a NON. Along these lines, I adapt the state-of-the-art bond percolation theory. Once the percolation threshold of a network becomes a local property, we can use the novel message passing theory to calculate the percolation threshold of each node. Classifying all the nodes according to this new measure I cam establish whether the network being studied should be cast within the framework of a NON or whether instead it is best represented by a more traditional description in terms of communities.

I further analyse real cases in which my method reveals a modular structure and which agrees with the condition of undergoing multiple percolation phase transitions. Moreover, I show that this modular structure implies the localization of the principal eigenvectors of the non-backtracking matrix. In such situations, the critical point of each module corresponds to an eigenvalue, renewing interest in the study of the spectral properties of the non-backtracking matrix.

Conclusions

From these results, I extract the following general conclusions.

First, I find that, due to topological constrains, considerable heterogeneity can explain part of the emergence of the high levels of clustering in real networks. At the same time, this indicates that the tree-like assumption may introduce major inaccuracies in random strongly heterogeneous networks.

Second, I show that triangles in real networks are distributed in a random fashion, in agreement with the perception that real complex networks are a product of a self-organized process in which edges are just a result of local interactions between nodes. This has an impact on the study of clustering in network processes, since it casts doubt on previous results derived from clustered network models in which triangles were organized in a very specific way.

Third, clustering makes weakly heterogeneous networks more fragile with respect to random failure of their connections, and less prone to spread infective agents. However, clustering in strongly heterogeneous networks can induce core–periphery organization, in which the core and periphery percolate inde-

pendently. This phenomenon, namely multiple percolation transitions, had not been observed previously. In this situation, clustering makes the core more robust and the periphery more fragile.

Furthermore, I analytically prove that such multiple percolation transitions are possible in networks that are sufficiently weakly connected. This new scenario has very important implications for different aspects of the analysis of the percolation properties of complex networks. On the one hand, the existence of multiple critical points changes the way we need to address percolation as a critical phenomenon. We should not develop theories to find the true and unique percolation threshold, but to reveal the set of critical points and the nodes involved in each one of them [70].

On the other hand, this new phenomenon implies that previous empirical methods for finding the percolation threshold are obsolete. The obvious incapacity to perform finite size scaling in a real finite system, together with the existence of multiple transitions, implies that no existent empirical method can be used to measure percolation thresholds.

Multiple percolation transitions also have a direct implication for the dynamics of epidemics. Previous conceptions assumed that the epidemic threshold depends on macroscopic properties of the network. Now, the epidemic threshold becomes a local property of the network. Therefore, there is the possibility that a disease may become endemic only in a finite fraction of the network, while the rest of the network remains healthy. This implies that, in contrast to previous beliefs, the location of an outbreak can also determine whether a disease will die out or become endemic. This finding could have important applications in the fields of vaccination and marketing strategies.

CHAPTER 9

List of publications

Related publications

- **Clustering of random scale-free networks**
  Pol Colomer-de-Simón & Marián Boguñá
  Physical Review E, 86, 026120, August 2012

- **Deciphering the global organization of clustering in real complex networks**
  Pol Colomer-de-Simón, M. Ángeles Serrano, Mariano G. Beiró, J. Ignacio Alvarez-Hamelin & Marián Boguñá
  Scientific Reports, 3:2517, August 2013

- **Double percolation phase transition in clustered complex networks**
  Pol Colomer-de-Simón & Marián Boguñá
  Physical Review X, 4(4):041020, October 2014

- **Local percolation thresholds**
  Pol Colomer-de Simón, Ali Faqeeh, James P. Gleeson & Marián Boguñá
  Work in preparation

Other publications

- **Quantifying randomness in real networks**
  Chiara Orsini, Marija M. Dankulov, Pol Colomer-de-Simón, Almerima Jamakovic, Priya Mahadevan, Amin Vahdat, Kevin E. Bassler, Zoltán Toroczkai, Marián Boguñá, Guido Caldarelli, Santo Fortunato & Dmitri Krioukov
  Nature Communications, 6:8627, October 2015

- **Emergence of coexisting percolating clusters in networks**
  Ali Faqeeh, Sergey Melnik, Pol Colomer-de-Simón & James P. Gleeson
  Arxiv, 1508.05590 August 2015

Code

- **RandNetGen**
  Pol Colomer-de Simón
  http://polcolomer.github.io/RandNetGen

- **LaNet-Vi 3.0**
  J. I Alvarez-Hamelin, Mariano G Beiró, and Pol Colomer-de Simón
  http://sourceforge.net/projects/lanet-vi

APPENDIX A

Appendix

A.1 Real networks data sets

In this section we give a description of the real world networks data sets that we have used throughout this thesis.

A.1.1 Internet AS

We use the autonomous system (AS) Internet topology of June 2009 extracted from data collected by the archipelago active measurement infrastructure developed by the Cooperative Association for Internet Data Analysis [28, 44]. The AS topology contains 23752 ASs and 58416 AS links, yielding the average AS degree $\bar{k} = 4.92$, clustering coefficient $\bar{c} = 0.51$ and maximum degree $k_{max} = 2778$.

A.1.2 Pretty-Good-Privacy netwok

Pretty-Good-Privacy (PGP) is the most popular encryptor algorithm aimed to maintain privacy in communication between peers in internet. This algorithm makes use of a pair of keys, one of them to encrypt the message, and its counterpart to decrypt the message. Both keys are generated in such a way that it is computationally infeasible to deduce one key from the other. Provided that everyone can generate a PGP key by himself, if anybody wants to know if a given key belongs really to the person stated in the key, he has to verify that. Hence exists a "signing procedure" where a person signs the public key of another, meaning that she trusts that the other person is who she claims to be. This procedure generates a web of peers that have signed public keys of another based on trust, and this is the so- called web of trust of PGP [25].
Here, we analyse the web of trust as it was on July 2001, when it comprised 191.548 keys and 286.290 signatures. Since we are mainly interested in the social character of the web of trust we only consider bidirectional signatures, i.e., peers who have mutually signed their keys. This filtering process guarantees mutual knowledge between connected peers and makes the PGP network a reliable proxy of the underlying social network. After the filtering process, we are left with an undirected network of 57.243 vertices, 61.837 edges, average degree $\bar{k} = 2.16$, clustering coefficient $\bar{c} = 0.50$ and maximum degree $k_{max} = 205$.

A.1.3 Escherichia coli's metabolism network

A simple abstraction of a given metabolism is given by its bipartite network representation. This amounts to consider metabolites and reactions as belonging to different subsets of nodes, with metabolites (irrespectively considered as reactants and products) linked to all reactions they take part in, thus avoiding connections between nodes of the same kind. Our network data set is the on-mode projection of the metabolism bipartite network of the bacteria Escherichia coli [158]. So nodes accounts for metabolites that are connected whenever they participate in the same reaction.
The resulting network has 1.010 nodes, 3.286 edges, average degree $\bar{k} = 6.51$, clustering coefficient $\bar{c} = 0.48$ and maximum degree $k_{max} = 54$.

A.1.4 Western US power grid network

This power grid dataset corresponds to an undirected, unweighted network representing the topology of the Western States Power Grid of the United States of America [170]. The resulting network has 4941 nodes, 6594 edges, an average degree of $\bar{k} = 2.67$, a clustering coefficient of $\bar{c} = 0.11$ and a maximum degree of $k_{max} = 19$. The k-core and m-core decompositions between the real power grid network, the clique based model, and maximally random models are shown in Fig. A.1 and the m-core visualizations at Fig. A.2.

A.1.5 US air transportation network

In the US air transportation network the nodes are airports and a link is the existence of a direct flight between two airports [157]. The network has 583 nodes, 1087s edges, an average degree of $\bar{k} = 3.73$, a clustering coefficient of $\bar{c} = 0.43$ and a maximum degree $k_{max} = 109$.

A.1.6 Human disease network

In the "human disease network" nodes represent disorders, and two disorders are connected to each other if they share at least one gene in which mutations are associated with both disorders [87]. The resulting network has 867 nodes, 1527 edges, an average degree of $\bar{k} = 3.52$, a clustering coefficient of $\bar{c} = 0.81$ and a maximum degree $k_{max} = 50$.

A.1.7 Pokec online social network

Pokec is one of the most popular on-line social network in Slovakia. Pokec has been provided for more than 10 years and connects more than 1.6 million people

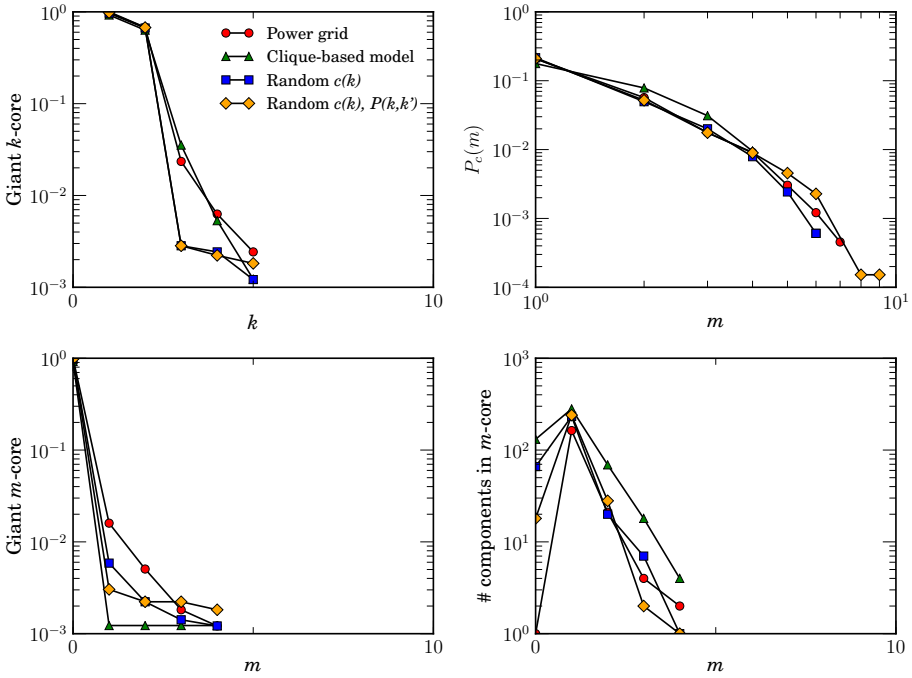


Figure A.1: Comparison of the k-core and m-core decompositions between the real power grid network, the clique based model, and maximally random models.

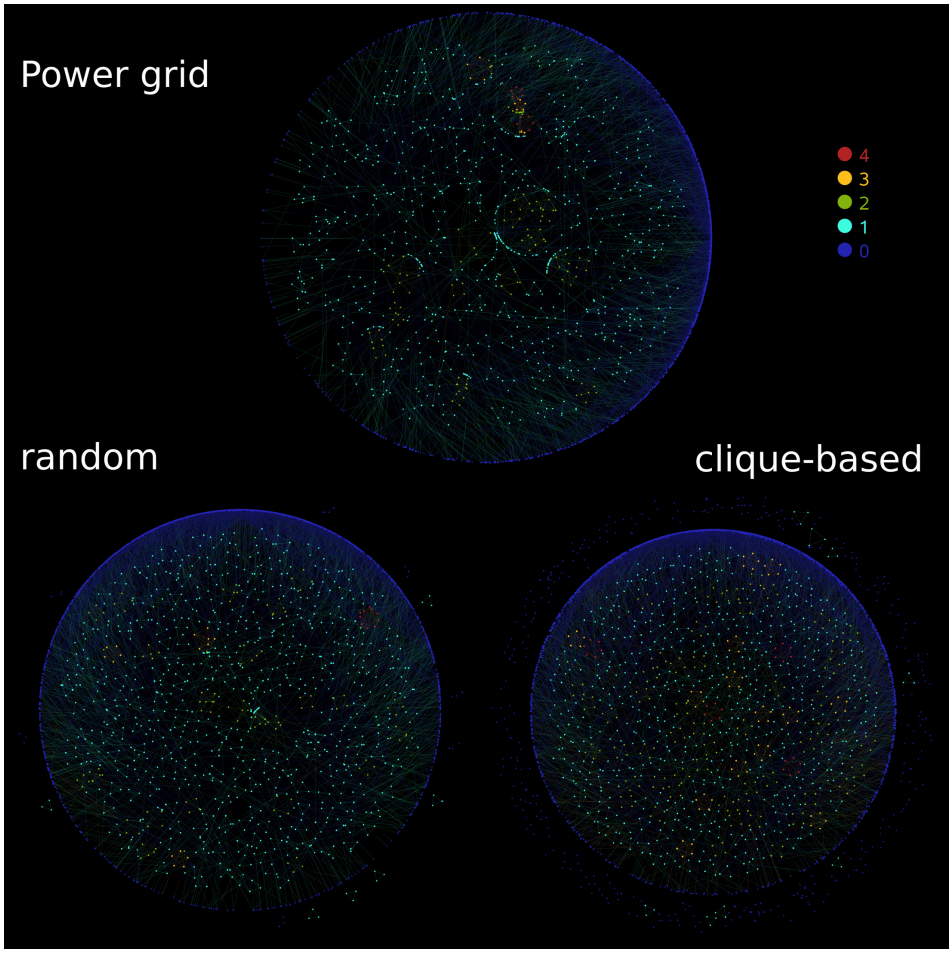


Figure A.2: m-core decomposition of the power grid and its random versions.

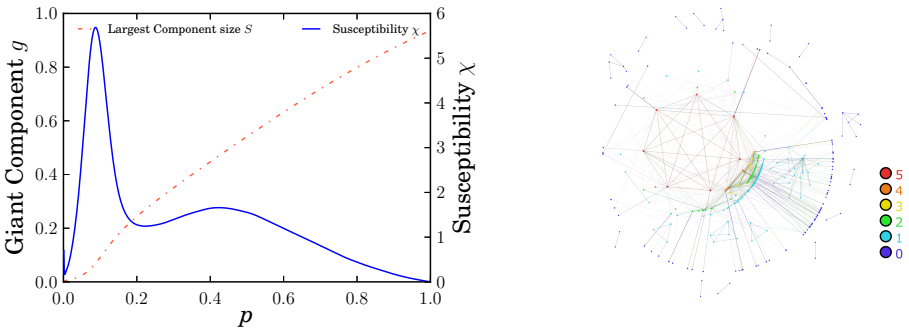


Figure A.3: Left: Bond percolation simulations for the US air transportation network. The relative size of the largest connected component g and its susceptibility χ as a function of the bond occupation probability p. Right:m-core decomposition

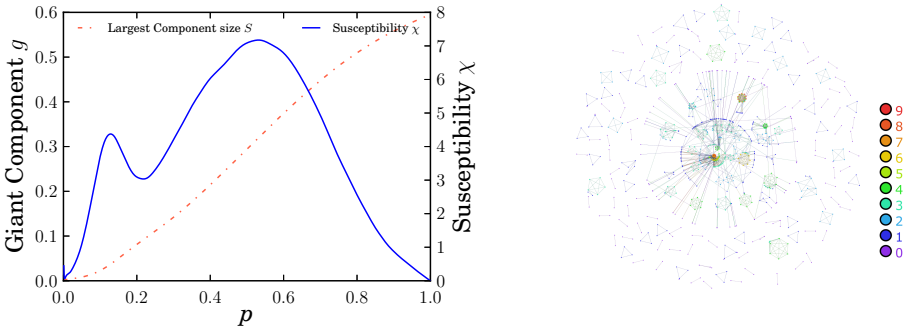


Figure A.4: Left: Bond percolation simulations for the human disease network network. The relative size of the largest connected component g and its susceptibility χ as a function of the bond occupation probability p. Right:m-core decomposition

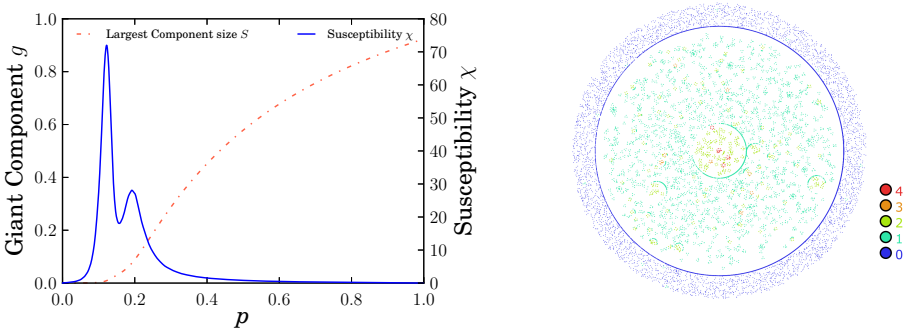


Figure A.5: Left: Bond percolation simulations for the Pokec On-line social network. The relative size of the largest connected component g and its susceptibility χ as a function of the bond occupation probability p. Right:m-core decomposition

by 2012. We analyse the undirected network by deleting all non-bidirectional links. For having a smaller system we only considered nodes that sign up into the on-line network before 2004. The resulting network has 44285 nodes, 75285 edges, an average degree of $\bar{k} = 3.4$, a clustering coefficient of $\bar{c} = 0.09$ and a maximum degree $k_{max} = 58$.

A.1.8 Gnutella pear-to-pear network

Gnutella was the first decentralized peer-to-peer file sharing network [110]. This data set is a snapshot of the Gnutella peer-to-peer file sharing network from August 9th in 2002. In this thesis we only consider the largest connected cluster that have 8104 nodes, 26008 edges, an average degree of $\bar{k} = 6.42$, a clustering

coefficient of $\bar{c} = 0.013$ and a maximum degree $k_{max} = 102$.

A.2 Mean-field critical exponents

Our exponent ν is the finite size scaling exponent in terms of the total number of nodes N and not of the one-dimensional scale L, as it is done in the book by Stauffer and Aharony [165]. This means that our critical exponent is just $\nu_{ours} = d\nu_{stauffer-aharony}$, where d is the dimension of the system. In the case of mean field, d must be replaced by the upper critical dimension, that for percolation is $d_u = 6$. This makes our mean field critical exponent $\nu_{our} = 3$. Perhaps it is more clear if we write the usual finite size scaling assumption for the size of the susceptibility. In terms of L it reads

$$\chi(L,p) = L^{\gamma/\nu} F\left[|p-p_c|L^{1/\nu}\right]$$

In terms of the total number of nodes, this is

$$\chi(N,p) = N^{\gamma/d\nu} F\left[|p-p_c|N^{1/d\nu}\right]$$

In the same book by Stauffer and Aharony it is proved that (Eq. (53), pg. 67)

$$d\nu = \frac{\tau-1}{\sigma}$$

and the mean field values of τ and σ are $\tau = 5/2$ $\sigma = 1/2$. This means that, again $d\nu = 3$.

Because we are normalizing the susceptibility according to equation 5.4 we are measuring the exponent $\gamma' = \gamma + \beta$ instead of γ. Therefore, if the mean field exponents are $\gamma = \beta = 1$ and $\nu = 3$, the exponent we are measuring in the scaling of the maximum of the susceptibility is $\gamma'/\nu = 2/3$.

APPENDIX B

Resum en català

Introducció

L'estudi de sistemes des del punt de vista de les xarxes és útil per concentrar-se en l'impacte que els patrons de interacció entre elements tenen en la funció de sistemes. La mesura i observació de xarxes reals revela que aquestes tenen unes estructures complexes, ni regulars ni totalment atzaroses, amb un efecte molt important en el seu comportament. A més a més, s'han trobat algunes propietats topològiques comunes entre diverses xarxes de naturalesa completament diferent. Aquest fet suggereix l'existència de patrons de formació comuns que determinen l'estructura i evolució de les xarxes.

Una de les propietats més comunes de les xarxes reals és l'alta presència de triangles o fort clustering. Al contrari que altres propietats topològiques, encara es desconeix l'origen de l'emergència del clustering i el seu efecte en l'estructura i funció del sistema. En primer lloc, això és degut a que la simple presència de triangles contradiu una hipòtesi molt utilitzada en la teoria de xarxes, complicant qualsevol possibilitat d'un tractament analític. En segon lloc, hi ha un manca de models de xarxes amb clustering apropiats que permetin un estudi empíric. Per tant, el clustering és un dels factors més importants que dificulta la possibilitat d'aplicar els resultats de la teoria de xarxes a casos reals.

En aquesta tesi estudiem el paper que juga el clustering en la estructura i funció de les xarxes complexes. En aquesta direcció, comencem estudiant el clustering generat pels models de xarxes més populars. Seguidament, mirem com es distribueixen els triangles en les xarxes reals. Finalment, ens concentrem en l'efecte del clustering en el clàssic problema de percolació d'enllaços. La nostra tria es basa en la relació que aquest procés simple té amb la robustesa i la dinàmica d'epidèmies en xarxes.

La percolació en xarxes amb clustering ha estat estudiat extensivament en anterioritat. Tot i això, els estudis anteriors són només vàlids per una estructura específica la qual mostrem que no reprodueix la organització global dels triangles present en les xarxes reals. Per tant, per respondre aquesta pregunta hem hagut primer de desenvolupar un model de xarxa amb clustering que reprodueixi la organització dels triangles de les xarxes reals. Finalment, hem fet servir el nostre model per estudiar com el clustering afecta a la posició del llindar de percolació en xarxes complexes.

Resultats

En el capítol 2, hem començat introduint els models de xarxa a l'atzar exponencials que generen xarxes al més a l'atzar possible donades unes restriccions, que són fixades en mitjana. D'aquesta col·lecció de models ens hem concentrat en les xarxes màxim a l'atzar amb un seqüència de graus esperats. Aquest model és la col·lecció canònica d'un dels models més popular en la teoria de xarxes, el model configurational. Fixant la seqüència de graus esperats enlloc dels graus exactes permet el tractament analític que és impossible en la col·lecció microcanònica.

En aquest marc, en el capítol 3 fem un estudi analític del clustering generat per xarxes a l'atzar sense escala, corregint resultats incorrectes anteriors per a xarxes molt heterogènies. Hem trobat que l'escalat correcte del coeficient de clustering de la col·lecció de xarxes a l'atzar sense escala amb exponent $2 < \gamma < 3$ és donat per l'equació. 3.11. Interessantment, per valors de l'exponent $\gamma \approx 2$, el clustering es manté quasi bé constant fins a valors extremadament grans de la mida de les xarxes però no és auto-promitjable. Aquest fet implica que les xarxes molt heterogènies poden tenir un nivell moderat de clustering només degut a les restriccions topològiques donades per la distribució de graus. Aquest fet contradiu la creença estesa de que les xarxes a l'atzar poden ser sempre aproximades com a xarxes d'arbre, violant unes les de les hipòtesis més comunes en la teoria de xarxes

Tot i això, el clustering generat en xarxes a l'atzar encara no és comparable a l'observat en moltes xarxes reals. Per tant, hi ha una necessitat de models de xarxes amb clustering que puguin ser usats per estudiar l'efecte del clustering en altres propietats topològiques i processos en xarxes. En aquest direcció, en la secció 2.5 hem desenvolupat un model que genera xarxes amb clustering d'una col·lecció exponencial a l'atzar via un procés de reconnexió esbiaixada.

El nostre model de xarxes amb clustering té dues característiques molt importants que el fan més convenient que altres models. Primer, és capaç de donar diferents nivells de clustering fixant a la vegada, la distribució de graus això com les correlacions de graus. Un fet important per tal de desacoblar els efectes del clustering d'aquestes dues propietats topològiques. Segon, el nostre model és un una xarxa exponencial, i per tant, maximitza l'entropia de la xarxa, fent les mínimes hipòtesis més que aquelles imposades per les restriccions. Per tant, en el nostre model, en contrast amb altre models de xarxes amb clustering, la distribució de triangles en la xarxa és la més a l'atzar possible.

A més a més, hem publicat el codi RandNetGen [48] que genera xarxes utilitzant el nostre model. El programa va més enllà del nostre model i pot generar qualsevol xarxa com a una xarxa a l'atzar exponencial utilitzant el mètode de reconnexió esbiaixada, utilitzant diferents propietats topològiques com a restric-

ció. El programa és agradable per a l'usuari i està publicat a la plataforma de codi obert colaborativa Github.

Més endavant, en el capítol 4 hem estudiat la distribució dels triangles en les xarxes reals. Per fer això, hem mirat l'estructura de l'm-core, que és molt més profunda si els triangles estan distribuïts a l'atzar que en una estructura modular, on hi ha correlacions en les multiplicitats de les connexions. A més a més, hem desenvolupat una eina de visualització, el LaNet-Vi 3.0 [16], que genera visualitzacions de l'estructura de l'm-core d'una xarxa. Els nostres resultats mostren que l'organització global del clustering en xarxes reals està molt més ben reproduïda per les nostres xarxes amb clustering màxim a l'atzar que per models de xarxes amb clustering anteriors, en els quals els triangles estan ordenats d'una manera molt específica. Per tant, el bon comportament del nostre model defineix el marc apropiat per estudiar com afecta el clustering en la percolació d'enllaços.

Després, hem utilitzat el nostre model per estudiar com afecta el clustering a les propietats de percolació de les xarxes. Amb aquest objectiu, hem comparat les propietats de la percolació d'enllaços de xarxes amb la mateixa seqüència de graus i correlacions de grau però diferents nivells de clustering.

Els nostres resultats mostren que els efectes del clustering depenen fortament de la heterogeneïtat de la xarxa. Per xarxes poc heterogènies el clustering incrementa el llindar de percolació fent-les més fràgils. Tot i això, per xarxes més heterogènies ($\gamma \leq 3.5$) un increment del clustering pot induir l'emergència d'una estructura de nucli perifèria. Aquesta organització redefineix l'espai de fases de la percolació de les xarxes complexes induint un nou fenomen, la doble transició de fase de percolació, en la qual el nucli i la perifèria percolen independentment l'una de l'altra.

En aquesta situació, incrementar el clustering fa disminuir el llindar de percolació del nucli i incrementa el llindar de percolació de la perifèria. Per xarxes més heterogènies, la mida del nucli no és macroscòpic, i per tant, el clustering només fa les xarxes més fràgils davant de la eliminació de connexions. Aquest fet reconcilia les dues interpretacions dominants de l'efecte del clustering en les propietats de percolació de les xarxes complexes.

A més a més, aquesta múltiple transició de fase de percolació que hem trobat redefineix completament la nostra comprensió prèvia del procés de percolació en xarxes complexes. En la teoria anterior, no era possible tenir dues o més transicions de fase consecutives associades al mateix trencament de simetria. Tot i això, en la secció 6.7, hem demostrat analíticament que aquestes transicions anòmales són en efecte possibles. Mostrem que dos mòduls macroscòpics dèbilment connectats poden percolar independentment mentre el seu intra-grau mitjà escali subliniarment amb la mida del sistema. Interessantment, aquest comportament també s'observa en una colecció de xarxes reals, una evidència

de la generalitat del fenomen.

Per tant, degut a la alta heterogeneïtat de les xarxes reals, ja no tenim un autèntic llindar de percolació, sinó un conjunt de punts crítics en els quals diferents parts de la xarxa percolen. Aquest fet canvia completament com hem d'afrontar el problema de la percolació i ens obliga a reinterpretar tots els esforços teòrics que intenten trobar un sol llindar de percolació.

La possibilitat de que una xarxa tingui diferents parts, o mòduls, que siguin independents en termes de processos de xarxes, suggereix la idea de considerar cada mòdul com a una xarxa i el sistema sencer com a una xarxa de xarxes. En aquesta direcció, hem readaptat la teoria més moderna de percolació [99]. Un cop el llindar de percolació d'una xarxa es torna una propietat local, podem utilitzar el tècnica de passar el missatge per calcular el llindar de percolació de cada node. Classificant tots els nodes d'acord a aquesta nova mesura hem sigut capaços de discernir si la xarxa sota estudi pot ser definida sota el marc de xarxa de xarxes, o si per el contrari, és millor representar-la amb la descripció més tradicional en termes de comunitats.

A més a més, hem analitzat casos reals en els quals el mètode revela una estructura modular que concorda amb la condició necessària per tenir una transició múltiple de percolació. Altrament, hem mostrat que aquesta estructura modular implica la localització del vector propi principal de la matriu non-backtracking. En aquesta situació, el punt crític de cada mòdul correspon a un valor propi, potenciant l'estudi de les propietats espectrals de la matriu non-backtracking.

Conclusions

D'aquests resultats que acabem de presentar podem extreure les següents conclusions generals.

Primer, hem trobat que, degut a restriccions topològiques, l'alta heterogeneïtat pot explicar part de la emergència dels alts nivells de clustering que trobem a les xarxes reals. Al mateix temps, aquest fet ens afecta a la precisió que la hipòtesi de tractar les xarxes com a arbres pot tenir en xarxes a l'atzar molt heterogènies.

Segon, hem ensenyat que en les xarxes reals els triangles estan distribuïts d'una manera atzarosa, d'acord amb la percepció que les xarxes complexes són un producte d'un procés auto-organitzat en el qual les connexions són resultat d'una interacció local entre nodes. Aquest fet té un impacte important en l'estudi de processos en xarxes amb clustering ja que posa dubtes sobre resultats previs derivats de models en els quals els triangles estan organitzats de manera molt específica.

Tercer, el clustering fa les xarxes poc heterogènies molt fràgils davant la fal-

lida de connexions a l'atzar i menys propenses a propagar agents infectats. Tot i això, el clustering en xarxes molt heterogènies poden induir una organització de nucli-perifèria en el qual el nucli i la perifèria percolen independentment. Aquest fenomen, una transició de percolació múltiple, no havia estat observat abans. En aquesta situació, el clustering fa el nucli més robust i la perifèria més fràgil.

Altrament, hem demostrat analíticament que una transició de percolació múltiple com aquesta és en efecte possible en xarxes dèbilment connectades. Aquest nou escenari té unes implicacions molt importants en diferents aspectes de l'anàlisi de les propietats de percolació de les xarxes complexes. Per una altra banda, l'existència de múltiples punts crítics canvia la manera en la qual hem d'abordar la percolació com a fenomen crític. Ja no hem de desenvolupar teories que trobin un verdader i únic llindar de percolació, sinó un conjunt de punt crítics i els nodes involucrats en cada una de ells [70].

Per altra banda, aquest fenomen fa que els mètodes empírics existents de trobar el llindar de percolació poc precisos. La òbvia incapacitat de fer una escalatge de mida finita en un sistema real finit i la existència de transicions múltiples implica que cap mètode empíric existent no pot ser utilitzat per mesurar el llindar de percolació.

La transició múltiple de percolació també té una implicació directa en dinàmica d'epidèmies. Les concepcions prèvies assumien que el llindar epidèmic depèn únicament en propietats macroscòpiques de la xarxa. Ara, el llindar epidèmic passar a ser una propietat local de la xarxa. Per tant, hi ha la possibilitat de que una malaltia infecciosa sigui pandèmica només en una fracció finita de la xarxa, mentre la resta de la xarxa estigui totalment sana. Aquest fet implica que, en contra de concepcions prèvies, l'origen d'un brot d'una malaltia infecciosa sigui determinant alhora de determinar si la malaltia esdevindrà pandèmica o no, amb importants aplicacions a estratègies de vacunació i de màrqueting.

References

[1] Albert, R., Jeong, H., and Barabasi, A. L. Internet: Diameter of the World-Wide Web. *Nature*, 401(6749):130–131, 1999. 3

[2] Amaral, L. A. N., Scala, A., Barthelemy, M., and Stanley, H. E. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–52, 2000. 4

[3] Anand, K. and Bianconi, G. Entropy measures for networks: Toward an information theory of complex topologies. *Physical Review E*, 80(4):1–4, 2009. 23

[4] Anand, K., Bianconi, G., and Severini, S. Shannon and von Neumann entropy of random networks with heterogeneous expected degree. *Physical Review E*, 83(3):1–8, 2011. 23

[5] Anderson, R. M. and May, R. M. *Infectious Diseases of Humans*. Oxford University Press, Oxford, 1991. 66

[6] Annibale, A., Coolen, A. C. C., Fernandez, L., Fraternali, F., and Kleinjung, J. Tailored graph ensembles as proxies or null models for real networks I: tools for quantifying structure. *Journal of Physics A*, 42(48):25, 2009. 22, 28, 29

[7] Arcangelis, L. D., Redner, S., and Coniglio, A. Anomalous voltage distribution of random networks and a new model for the backbone at the percolation threshold. *Physical Review B*, 31(7):4725, 1985. 59

[8] Bailey, N. *The mathematical theory of infectious diseases and its applications*. Hafner Press, New York, 1975. 66

[9] Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *Science*, 286(October):509–512, 1999. 13, 43

[10] Barabási, A.-L. and Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2):101–113, 2004. 5

[11] Barkema and Newman, M. E. J. *Monte Carlo Methods in Statistical Physics*. Clarendon Press, Oxford, 1999. 62

[12] Barrat, A. and Pastor-Satorras, R. Rate equation approach for correlations in growing network models. *Physical Review E*, 71(3):1–13, 2005. 38

[13] Barrat, A., Barthelemy, M., and Vespignani, A. *Dynamical Processes on Complex Networks*. Cambridge University Press, Cambridge, 2012. 4

[14] Baxter, G. J., Dorogovtsev, S. N., Goltsev, a. V., and Mendes, J. F. F. Avalanche collapse of interdependent networks. *Physical Review Letters*, 109(24): 248701, 2012. 71

[15] Beiró, M. G., Alvarez-Hamelin, J. I., and Busch, J. R. A low complexity visualization tool that helps to perform complex systems analysis. *New Journal of Physics*, 10(125003), 2008. 47, 48, 49

[16] Beiró, M. G., Colomer-de Simón, P., and Alvarez-Hamelin, J. I. LaNet-Vi 3.0, 2014. URL `http://sourceforge.net/projects/lanet-vi/`. 49, 105, 119

[17] Bender, E. a. and Canfield, E. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978. 14, 33

[18] Berger, N., Borgs, C., Chayes, J. T., and Saberi, A. On the spread of viruses on the internet. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '05, pages 301–310, Philadelphia, PA, USA, 2005. Society for Industrial and Applied Mathematics. ISBN 0-89871-585-7. 71

[19] Bianconi, G. and Dorogovtsev, S. N. Multiple percolation transitions in a configuration model of a network of networks. *Physical Review E*, 89(6): 062814, 2014. 72, 91

[20] Bianconi, G. and Marsili, M. Effect of degree correlations on the loop structure of scale-free networks. *Physical Review E*, 73(6):1–6, 2006. 18, 36

[21] Bianconi, G., Caldarelli, G., and Capocci, A. Loops structure of the Internet at the autonomous system level. *Physical Review E*, 71(6):11–14, 2005. 14, 34

[22] Boguñá, M. and Pastor-Satorras, R. Class of correlated random networks with hidden variables. *Physical Review E*, 68(3):036112, 2003. 8, 24, 25, 36

[23] Boguñá, M. and Serrano, M. Á. Generalized percolation in random directed networks. *Physical Review E*, 72(January):1–7, 2005. 68

[24] Boguñá, M., Pastor-Satorras, R., and Vespignani, A. Absence of epidemic threshold in scale-free networks with degree correlations. *Physical review letters*, 90(2):028701, 2003. 4, 18, 71

[25] Boguñá, M., Pastor-Satorras, R., Díaz-Guilera, A., and Arenas, A. Models of social networks based on social distance attachment. *Physical Review E*, 70(5 2):056122, 2004. 44, 111

[26] Boguñá, M., Pastor-Satorras, R., and Vespignani, A. Cut-offs and finite size effects in scale-free networks. *European Physical Journal B*, 38(2):205–209, 2004. 14, 33

[27] Boguñá, M., Krioukov, D., and Claffy, K. Navigability of Complex Networks. *Nature Physics*, 5(1):74–80, 2007. 18, 54

[28] Boguñá, M., Papadopoulos, F., and Krioukov, D. Sustaining the Internet with hyperbolic mapping. *Nature communications*, 1:62, 2010. 4, 18, 44, 53, 54, 111

[29] Boguñá, M., Castellano, C., and Pastor-Satorras, R. Nature of the epidemic threshold for the susceptible-infected-susceptible dynamics in networks. *Physical Review Letters*, 111(6):68701, 2013. 71

[30] Bollobás, B. The distribution of the maximum degree of a random graph. *Discrete Mathematics*, 32:201–203, 1980. 33

[31] Bollobás, B. A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs. *European Journal of Combinatorics*, 1(4): 311–316, 1980. 14

[32] Bollobás, B. *Random graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2001. ISBN 9780521797221. 4

[33] Bollobás, B., Borgs, C., Chayes, J. T., and Riordan, O. Percolation on dense graph sequences. *Annals of Probability*, 38(1):150–183, 2010. 69, 93

[34] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. Graph structure in the Web. *Computer Networks*, 33:309–320, 2000. 3

[35] Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., and Havlin, S. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291): 1025–1028, 2010. 71, 91

[36] Burda, Z. and Krzywicki, A. Uncorrelated random networks. *Physical Review E*, 67(4 Pt 2):046118, 2003. 33

[37] Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts, D. J. Network robustness and fragility: percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471, 2000. 15, 64, 69, 93

[38] Callaway, D. S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. J., and Strogatz, S. H. Are randomly grown graphs really random? *Physical Review E*, 64(4 Pt 1):041902, 2001. 71

[39] Catanzaro, M., Boguñá, M., and Pastor-Satorras, R. Generation of uncorrelated random scale-free networks 027103. *Physical Review E*, 71(2):027103, 2005. 25, 33, 34, 36, 37

[40] Černý, V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, 1985. 26

[41] Chatterjee, S. and Durrett, R. Contact processes on random graphs with power law degree distributions have critical value 0. *Annals of Probability*, 37(6):2332–2356, 2009. 71

[42] Chen, W., Cheng, X., Zheng, Z., Chung, N. N., D'Souza, R. M., and Nagler, J. Unstable supercritical discontinuous percolation transitions. *Physical Review E*, 88(4):042152, 2013. 72

[43] Chen, W., Nagler, J., Cheng, X., Jin, X., Shen, H., Zheng, Z., and D'Souza, R. M. Phase transitions in supercritical explosive percolation. *Physical Review E*, 87(5):052130, 2013. 72

[44] Claffy, K., Hyun, Y., Keys, K., Fomenkov, M., and Krioukov, D. Internet mapping: From art to science. *Proceedings - Cybersecurity Applications and Technology Conference for Homeland Security, CATCH 2009*, pages 205–211, 2009. 111

[45] Clauset, A., Shalizi, C. R., and Newman, M. E. J. Power-Law Distributions in Empirical Data. *Society for Industrial and Applied Mathematics*, 51(4): 661–703, 2009. 7

[46] Cohen, R., Erez, K., Ben-Avraham, D., and Havlin, S. Resilience of the Internet to random breakdowns. *Physical Review Letters*, 85(21):4626–4628, 2000. 15, 69, 93

[47] Cohen, R., Ben-Avraham, D., and Havlin, S. Percolation critical exponents in scale-free networks. *Physical Review E*, 66(3):036113, 2002. 62

[48] Colomer-de Simón, P. RandNetGen, 2014. URL `http://polcolomer.github.io/RandNetGen`. 104, 118

[49] Colomer-de Simón, P. and Boguñá, M. Double percolation phase transition in clustered complex networks. *Physical Review X*, 4:041020, 2014. 92, 94

[50] Colomer-de Simón, P., Serrano, M. Á., Beiró, M. G., Alvarez-Hamelin, J. I., and Boguñá, M. Deciphering the global organization of clustering in real complex networks. *Scientific reports*, 3:2517, 2013. 44

[51] Coolen, A. C. C., Martino, A., and Annibale, A. Constrained Markovian Dynamics of Random Graphs. *Journal of Statistical Physics*, 136(6):1035–1067, 2009. 28

[52] Costa, L. D. F., Rodrigues, F. a., Travieso, G., and Boas, P. R. V. Characterization of complex networks: A survey of measurements. *Advance in Physics*, 56(June 2007):78, 2005. 8

[53] Cover, T. M. and Thomas, J. A. *Elements Of Information Theory Notes*. John Wiley, New York, 2006. ISBN 0471062596. 23

[54] Csermely, P., London, A., Wu, L.-Y., and Uzzi, B. Structure and dynamics of core/periphery networks. *Journal of Complex Networks*, 1(2):93–123, 2013. 71

[55] Czabarka, É., Dutle, A., Erdős, P. L., and Miklós, I. On realizations of a joint degree matrix. *Discrete Applied Mathematics*, 181:283–288, 2015. 27

[56] Danon, L., Duch, J., Diaz-Guilera, A., and Arenas, A. Comparing community structure identification. *Journal of Statistical Mechanics*, 09008:10, 2005. 11

[57] de Sola Pool, I. and Kochen, M. Contacts and influence. *Social Networks*, 1:5–51, 1978. 2

[58] de Solla Price, D. Networks of scientific papers. *Science*, 149:510–515, 1965. 4

[59] de Solla Price, D. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976. 13

[60] Del Genio, C. I., Kim, H., Toroczkai, Z., and Bassler, K. E. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PLoS ONE*, 5(4):e10012, 2010. 35

[61] Del Genio, C. I., Gross, T., and Bassler, K. E. All scale-free networks are sparse. *Physical Review Letters*, 107(17):1–4, 2011. 33

[62] Dorogovtsev, S. N. Clustering of correlated networks. *Physical Review E*, 69(2 2):027104, 2004. 9

[63] Dorogovtsev, S. N. and Mendes, J. F. F. *Evolution of Networks: From Biological Nets to the Internet and WWW*, volume 57. Oxford University Press, Oxford, 2003. ISBN 0198515901. 1, 38, 43

[64] Dorogovtsev, S. N., Mendes, J. F. F., and Samukhin, a. N. Anomalous percolation properties of growing networks. *Physical Review E*, 64(6 Pt 2): 066110, 2001. 71

[65] Dorogovtsev, S. N., Goltsev, a. V., and Mendes, J. F. F. K-core organization of complex networks. *Physical Review Letters*, 96(4):040601, 2006. 46

[66] Erdős, P. and Rényi, A. On random graphs. *Publ Math*, 6:290–297, 1959. 2, 12

[67] Erdős, P. and Rényi, A. On the evolution of random graphs. *Publications Mathematics Institute Hungarian Academy of Science*, 5:17, 1960. 12

[68] Euler, L. and Euler, L. Solutio problematis ad geometrian situs pertinentis. *Comentarii academiae scientarum Petropolitanae*, 8:128–140, 1736. 2

[69] Faloutsos, M., Faloutsos, P., and Faloutsos, C. On Power-law relationship of the internet topology. In *SIGCOMM'99*, volume 53, pages 1689–1699, 1999. ISBN 9788578110796. 3, 4

[70] Faqeeh, A., Melnik, S., Colomer-de Simón, P., and Gleeson, J. P. Emergence of coexisting percolating clusters in networks. *Arxiv*, 1508.05590(1), 2015. 99, 107, 121

[71] Faqeeh, A., Melnik, S., and Gleeson, J. P. Network cloning unfolds the effect of clustering on dynamical processes. *Physical Review E*, 91(5):052807, 2015. 17, 19, 69

[72] Fortunato, S. Community detection in graphs. *Physics Reports*, 486(3-5): 75–174, 2010. 4

[73] Foster, D. V., Foster, J. G., Paczuski, M., and Grassberger, P. Communities, clustering phase transitions, and hysteresis: Pitfalls in constructing network ensembles. *Physical Review E*, 81(4):046115, 2010. 45, 78

[74] Foster, D. V., Foster, J. G., Grassberger, P., and Paczuski, M. Clustering drives assortativity and community structure in ensembles of networks. *Physical Review E*, 84(6):066117, 2011. 53

[75] Frank, K. T., Petrie, B., Choi, J. S., and Leggett, W. C. Trophic cascades in a formerly cod-dominated ecosystem. *Science*, 308(5728):1621–1623, 2005. 1

[76] Frank, O. and Strauss, D. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986. 45

[77] Fronczak, A., Fronczak, P., and Hołyst, J. a. Mean-field theory for clustering coefficients in Barabási-Albert networks. *Physical Review E*, 68(4 Pt 2): 046126, 2003. 13

[78] Gallos, L. K., Song, C., and Makse, H. a. Scaling of degree correlations and its influence on diffusion in scale-free networks. *Physical Review Letters*, 100(24):248701, 2008. 18

[79] Gao, J., Buldyrev, S. V., Havlin, S., and Stanley, H. E. Robustness of a network of networks. *Physical Review Letters*, 107(19):1–5, 2011. 91

[80] Garlaschelli, D. and Loffredo, M. I. Maximum likelihood: Extracting unbiased information from complex networks. *Physical Review E*, 78(1):1–4, 2008. 23

[81] Giles, J. Making the Links. *Nature*, 488:448–450, 2012. 1

[82] Gjoka, M., Kurant, M., and Irvine, U. C. 2 . 5K-Graphs : from Sampling to Generation. In *IEEE NetSciCom '14*, 2014. 28

[83] Gleeson, J. P. Bond percolation on a class of clustered random networks. *Physical Review E*, 80(3):405005, 2009. 43, 70

[84] Gleeson, J. P. Bond percolation on a class of clustered random networks. *Physical Review E*, 80(3):036107, 2009. 15, 19, 43, 44, 70, 71, 90

[85] Gleeson, J. P. and Melnik, S. Analytical results for bond percolation and k -core sizes on clustered networks. *Physical Review E*, 80(4):046121, 2009. 19, 43, 70

[86] Gleeson, J. P., Melnik, S., and Hackett, A. How clustering affects the bond percolation threshold in complex networks. *Physical Review E*, 81(6): 066114, 2010. 19, 43, 70, 71, 73, 90

[87] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007. 98, 112

[88] Goltsev, a. V., Dorogovtsev, S. N., Oliveira, J. G., and Mendes, J. F. F. Localization and spreading of diseases in complex networks. *Physical Review Letters*, 109(12):128702, 2012. 93

[89] Gradshteyn, I. S. *Tables of Integrals, Series, and Products*. Academic Press, San Diego, &th edition, 1988. ISBN 0122947576. 36

[90] Grassberger, P. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63(2):157–172, 1983. 67

[91] Gregori, E., Lenzini, L., and Orsini, C. K-Dense communities in the Internet AS-level topology graph. *Computer Networks*, 57(1):213–227, 2013. 47

[92] Hastings, W. K. and Apr, N. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. 26

[93] Henley, C. L. Statics of a "self-organized" percolation model, 1993. 59

[94] Hethcote, H. W. The mathematics of infectious diseases. *SIAM Review*, 42 (4):599–653, 2000. 66

[95] Jamakovic, A., Mahadevan, P., Vahdat, A., Boguñá, M., and Krioukov, D. How small are building blocks of complex networks. *Arxiv*, 0908.1143, 2009. 53

[96] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, a. L. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000. 3, 4

[97] Jovanović, B., Buldyrev, S. V., Havlin, S., and Stanley, H. E. Punctuated equilibrium and history-dependent percolation. *Physical Review E*, 50(4): 2403–2406, 1994. 59

[98] Karrer, B. and Newman, M. E. J. Random graphs containing arbitrary distributions of subgraphs. *Physical Review E*, 82(6):066118, 2010. 15, 43, 70

[99] Karrer, B., Newman, M. E. J., and Zdeborová, L. Percolation on sparse networks. *Physical Review Letters*, 113(20):1–5, 2014. 17, 18, 69, 93, 94, 95, 98, 100, 120

[100] Keeling, M. J. The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society B*, 266(1421):859–867, 1999. 19

[101] Kempe, D., Kleinberg, J. M., and Tardos, É. Maximizing the spread of influence through a social network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*, page 137, 2003. 4

[102] Kenah, E. and Robins, J. M. Second look at the spread of epidemics on networks. *Physical Review E*, 76(3):036113, 2007. 68

[103] Kirkpatrick, S. Optimization by Simulated. *Science*, 220(4598), 1983. 26

[104] Kiss, I. Z. and Green, D. M. Comment on "properties of highly clustered networks". *Physical Review E*, 78(4):048101, 2008. 19, 71, 73, 90

[105] Klein-Hennig, H. and Hartmann, A. K. Bias in generation of random graphs. *Physical Review E*, 85(2):1–7, 2012. 33

[106] Klemm, K. and Eguíluz, V. M. Growing scale-free networks with small-world behavior. *Physical Review E*, 65(5):1–4, 2002. 13

[107] Krapivsky, P. L., Redner, S., and Leyvraz, F. Connectivity of growing random networks. *Physical Review Letters*, 85(21):4629–4632, 2000. 13

[108] Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguñá, M. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010. 15, 54

[109] Krzakala, F. and Moore, C. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–40, 2013. 101

[110] Leskovec, J., Kleinberg, J. M., and Faloutsos, C. Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2006. 97, 115

[111] Liljeros, F., Edling, C. R., Amaral, L. a., Stanley, H. E., and Aberg, Y. The web of human sexual contacts. *Nature*, 411(6840):907–908, 2001. 1

[112] Lloyd, a. L. and May, R. M. Epidemiology. How viruses spread among computers and people. *Science*, 292(5520):1316–1317, 2001. 71

[113] Maas, G. a., Bial, M., Fijalkowski, J., and Ucte. System disturbance on 4 November 2006. Technical Report November, UCTE, 2007. 2

[114] Marro, J. and Dickman, R. *Nonequilibrium phase transitions in lattice models*. Cambridge University Press, Cambridge, 1999. ISBN 9780521480628. 62

[115] Maslov, S. and Sneppen, K. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002. 4

[116] Maslov, S., Sneppen, K., and Zaliznyak, A. Detection of topological patterns in complex networks: Correlation profile of the internet. *Physica A*, 333(1-4):529–540, 2004. 8, 14

[117] Melnik, S., Hackett, A., Porter, M. A., Mucha, P. J., and Gleeson, J. P. The unreasonable effectiveness of tree-based theory for networks with clustering. *Physical Review E*, 83(3):036112, 2011. 19, 69

[118] Melnik, S., Porter, M. A., Mucha, P. J., and Gleeson, J. P. Dynamics on modular networks with heterogeneous correlations. *Chaos*, 24(2):023106, 2014. 84

[119] Metropolis, N. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087, 1953. 26

[120] Meyer, D., Zhang, L., and Fall, K. Report from the IAB Workshop on Routing and Addressing. Technical report, IAB Workshop on Routing and Addressing, 2007. 16

[121] Miller, J. C. Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. *Physical Review E*, 76(1):010101, 2007. 68

[122] Miller, J. C. Percolation and epidemics in random clustered networks. *Physical Review E*, 80(2 Pt 1):020901, 2009. 19, 43, 70, 71, 73, 90

[123] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. 4, 9, 45

[124] Milo, R., Kashtan, N., Itzkovitz, S., Newman, M. E. J., and Alon, U. On the uniform generation of random graphs with prescribed degree sequences. *Arxiv*, 0312028v2:1–4, 2004. 29

[125] Mollison, D. Spatial Contact Models for Ecological and Epidemic Spread. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(3): 283–326, 1977. 67

[126] Molloy, M. and Reed, B. A critical point for random graphs with a given sequence degree. *Random Structures Algorithms*, 6(2-3):161–179, 1995. 33

[127] Molloy, M. and Reed, B. The Size of the Giant Component of a Random Graph with a Given Degree Sequence. *Combinatorics, Probability and Computing*, 7(03):29–305, 2000. 33

[128] Moore, C. and Newman, M. E. J. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5 Pt B):5678–5682, 2000. 59

[129] Nagler, J., Tiessen, T., and Gutch, H. W. Continuous percolation with discontinuities. *Physical Review X*, 2(3):031009, 2012. 72

[130] Newman, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–9, 2001. 3

[131] Newman, M. E. J. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2 Pt 2):025102, 2001. 9

[132] Newman, M. E. J. Properties of highly clustered networks. *Physical Review E*, 68(2 Pt 2):026121, 2003. 19, 71, 90

[133] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):1–19, 2006. 11, 99

[134] Newman, M. E. J. Random Graphs with Clustering. *Physical Review Letters*, 103(5):058701, 2009. 15, 19, 43, 70, 71, 73, 90

[135] Newman, M. E. J. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010. ISBN 0199206651, 9780199206650. 4, 9, 14, 33, 34, 43

[136] Newman, M. E. J. and Girvan, M. Finding and evaluating community structure in networks. *Physical Review E*, 69(2 2):1–15, 2004. 11

[137] Newman, M. E. J. and Ziff, R. M. Efficient Monte Carlo algorithm and high-precision results for percolation. *Physical Review Letters*, 85(19):4104–4107, 2000. 60, 73, 92

[138] Newman, M. E. J., Barabási, A.-L., and Watts, D. J. *The structure and dynamics of networks*. Princeton University Press, Princeton, New Jersey, 2006. 2

[139] Orsini, C., Gregori, E., Lenzini, L., and Krioukov, D. Evolution of the Internet k-Dense Structure. *IEEE/ACM Transactions on Networking*, 22(6):1–12, 2013. 47

[140] Orsini, C., Dankulov, M. M., Colomer-de Simón, P., Jamakovic, A., Mahadevan, P., Vahdat, A., Bassler, K. E., Toroczkai, Z., Boguñá, M., Caldarelli, G., Fortunato, S., and Krioukov, D. Quantifying randomness in real networks. *Nature Communications*, 6(May):8627, 2015. 31, 72

[141] Papadopoulos, F., Boguñá, M., and Krioukov, D. Popularity versus Similarity in Growing Networks. *Nature*, 489:537–540, 2011. 54

[142] Park, J. and Newman, M. E. J. Origin of degree correlations in the Internet and other networks. *Physical Review E*, 68(2 Pt 2):026112, 2003. 14, 18, 25, 33

[143] Park, J. and Newman, M. E. J. Solution for the properties of a clustered network. *Physical Review E*, 72(2):026136, 2005. 78

[144] Pastor-Satorras, R. and Vespignani, A. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001. 15, 71

[145] Pastor-Satorras, R. and Vespignani, A. Immunization of complex networks. *Physical Review E*, 65(3):1–8, 2002. 4

[146] Pastor-Satorras, R. and Vespignani, A. *Evolution and Structure of the Internet*. Cambridge University Press, 2004. 1

[147] Phil, A. *Thermodynamics and Statistical Mechanics: Equilibirium by Entropy Maximisation*. Academic Press, London, 2002. ISBN 9780120663217. 23

[148] Radicchi, F. Predicting percolation thresholds in networks. *Physical Review E*, 010801:1–5, 2015. 62, 69, 92, 93

[149] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658–2663, 2004. 47

[150] Ravasz, E. and Barabási, A.-L. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, 2003. 9

[151] Roberts, E. S. and Coolen, A. C. C. Unbiased degree-preserving randomization of directed binary networks. *Physical Review E*, 85(4):046103, 2012. 28

[152] Saito, K., Yamada, T., and Kazama, K. Extracting communities from complex networks by the k-dense method. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E91-A(11): 3304–3311, 2008. 47

[153] Serrano, M. Á. and Boguñá, M. Percolation and epidemic thresholds in clustered networks. *Physical Review Letters*, 97(8):088701, 2006. 19, 43

[154] Serrano, M. Á. and Boguñá, M. Clustering in complex networks. II. Percolation properties. *Physical Review E*, 74(5):056115, 2006. 19, 43, 47, 71, 72, 90

[155] Serrano, M. Á. and Boguñá, M. Clustering in complex networks. I. General formalism. *Physical Review E*, 74(5):056114, 2006. 19, 34, 47, 73

[156] Serrano, M. Á., Krioukov, D., and Boguñá, M. Self-similarity of complex networks and hidden metric spaces. *Physical Review Letters*, 100(7): 078701, 2008. 15, 25, 35, 54

[157] Serrano, M. Á., Boguñá, M., and Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009. 112

[158] Serrano, M. Á., Boguñá, M., and Sagués, F. Uncovering the hidden geometry behind metabolic networks. *Molecular bioSystems*, 8(3):843–50, 2011. 4, 18, 44, 54, 112

[159] Serrano, M. Á., Krioukov, D., and Boguñá, M. Percolation in self-similar networks. *Physical Review Letters*, 106(4):048701, 2011. 25, 38

[160] Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3, 2001. 23, 27

[161] Solomon, S., Weisbuch, G., De Arcangelis, L., Jan, N., and Stauffer, D. Social percolation models. *Physica A*, 277(1):239–247, 2000. 59

[162] Solomonoff, R. An exact method for the computation of the connectivity of random nets. *The Bulletin of Mathematical Biophysics*, 14(2):153–157, 1952. 12

[163] Solomonoff, R. and Rapoport, A. Connectivity of random nets. *The Bulletin of Mathematical Biophysics*, 13(2):107–117, 1951. 2, 12

[164] Son, S.-W., Bizhani, G., Christensen, C., Grassberger, P., and Paczuski, M. Percolation theory on interdependent networks based on epidemic spreading. *Europhysics Letters*, 97(1):16006, 2012. 71

[165] Stauffer, D. and Aharony, A. *Introduction to Percolation Theory*. Taylor & Francis, London, 1985. ISBN 0748402535. 4, 17, 59, 60, 61, 62, 64, 77, 86, 92, 96, 116

[166] Strauss, D. On a General Class of Models for Interaction. *SIAM Review*, 28 (4):513–527, 1986. 78

[167] Trapman, P. On analytical approaches to epidemics on networks. *Theoretical Population Biology*, 71(2):160–173, 2007. 43, 70

[168] Travers, J. and Milgram, S. An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443, 1969. 3, 9

[169] Vázquez, A. and Moreno, Y. Resilience to damage of graphs with degree correlations. *Physical Review E*, 67(1 Pt 2):015101, 2003. 18

[170] Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. 1, 3, 9, 15, 43, 112

[171] Zlatic, V., Garlaschelli, D., and Caldarelli, G. Complex networks with arbitrary edge multiplicities. *Europhysics Letters*, 97(2):28005, 2011. 47