Geometric Graphs, the Cosmic Web and Hypergraphs

by Bruno Coutinho

B.S. in Physics, University of Aveiro
M.S. in Physics, University of Aveiro

A dissertation submitted to

The Faculty of
the College of Science of
Northeastern University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

July 14, 2016

Dissertation directed by
Albert-László Barabási
Distinguished University Professor
Dedication

I dedicate this thesis to my fiancée Cristina and my sister Ana Raquel.
Acknowledgements

I wish to thank those who have supported me during these years. First, I would like to thank my advisor, Professor Albert-László Barabási for all time, support and direction he provided during these years. Second, I want to thank all the members of my dissertation committee for their support.

This thesis could not be completed without the help of Professor Yang-Yu Liu, Professor Arjun Dey, Dr. Sungryong Hong and Professor Hai-Jun Zhou. Thank you for all the discussions, ideas, time, help and inspiration you provided. Special thanks also go to James Stanfill for all the help editing my thesis, and Suzanne Aleva for administrative work. I also want to use this opportunity to thank everyone at the Center for Complex Network Research lab who discussed new ideas with me.

Finally, I wish to thank FCT for the financial support provided through grant no. SFRH/BD/79723/2011.
Abstract of Dissertation


First we show how network science can be applied to study the cosmic web and we explore a hypothetical spreading process at the cosmic level using concepts from network science. The concept of the cosmic web, viewing the Universe as a set of discrete galaxies held together by gravity, is deeply engrained in cosmology. Yet, little is known about the most effective construction and the characteristics of the underlying network. We ex-
plore seven network construction algorithms that use various galaxy properties, from their location to their size and relative velocity. We find that a model relying only on spatial proximity offers the best correlations between the physical characteristics of the connected galaxies. We show that the properties of the networks generated from simulations and observations are identical, unveiling a deep universality of the cosmic web. Second, we explore a generalization of graphs, called hypergraphs, which offer a much more faithful representation of many complex systems. We find that in most real-world hypergraphs two key combinatorial problems, the edge and the vertex cover problems, can be solved in polynomial time.
Table of Contents

1 Dedication ii
2 Acknowledgements iii
3 Abstract of Dissertation iv
4 Introduction 1
   4.1 Thesis structure and content 2
   4.2 Definitions and nomenclature 5
   4.3 Classic random network models 7
      4.3.1 Erdős-Rényi model 8
      4.3.2 Barabási-Albert model 10
   4.4 Network epidemics 11
   4.5 Role of topology in the complexity of combinatorial problems 13
   4.6 Large scale cosmic simulations 19
5 Spatial Networks 21
   5.1 Random geometric graph 22
   5.2 random k-nearest neighbor graph 26
   5.3 Moments of the degree distribution for k=1 29
6 The Network Behind the Cosmic Web 39
   6.1 Simulated and observational galaxy samples 40
   6.2 Models 43
   6.3 Pairwise correlations 45
   6.4 Randomized networks 49
   6.5 M3 applied to real data 52
   6.6 Conclusion 56
7 Spreading Across the Universe 57
   7.1 Model 58
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2</td>
<td>Traveling time</td>
<td>59</td>
</tr>
<tr>
<td>7.3</td>
<td>Linking length in a expanding box</td>
<td>63</td>
</tr>
<tr>
<td>7.4</td>
<td>Results</td>
<td>66</td>
</tr>
<tr>
<td>7.5</td>
<td>Discussion</td>
<td>67</td>
</tr>
<tr>
<td>8</td>
<td>Percolations on Hypergraphs</td>
<td>69</td>
</tr>
<tr>
<td>8.1</td>
<td>Giant component</td>
<td>72</td>
</tr>
<tr>
<td>8.2</td>
<td>$(K, S)$-core</td>
<td>75</td>
</tr>
<tr>
<td>8.2.1</td>
<td>$K = 2$ and $S = S_{\text{max}}$</td>
<td>76</td>
</tr>
<tr>
<td>8.2.2</td>
<td>$K = 2$ and $S = 2$</td>
<td>79</td>
</tr>
<tr>
<td>8.3</td>
<td>Discussion</td>
<td>81</td>
</tr>
<tr>
<td>8.A</td>
<td>Critical exponents and critical points</td>
<td>81</td>
</tr>
<tr>
<td>8.A.1</td>
<td>Emergence of the giant component in hypergraphs - critical point</td>
<td>81</td>
</tr>
<tr>
<td>8.A.2</td>
<td>Emergence of the giant component in hypergraphs - critical exponent</td>
<td>82</td>
</tr>
<tr>
<td>8.A.3</td>
<td>$(K, S)$-core</td>
<td>84</td>
</tr>
<tr>
<td>8.A.4</td>
<td>$(2, S_{\text{max}})$-core - critical point</td>
<td>84</td>
</tr>
<tr>
<td>8.A.5</td>
<td>$(2, \text{max}(r, 2))$-core - critical exponent</td>
<td>85</td>
</tr>
<tr>
<td>8.A.6</td>
<td>$(2, 2)$-core - critical point</td>
<td>88</td>
</tr>
<tr>
<td>8.A.7</td>
<td>$2, 2$-core - critical exponent</td>
<td>88</td>
</tr>
<tr>
<td>8.A.8</td>
<td>Simulations</td>
<td>89</td>
</tr>
<tr>
<td>9</td>
<td>Greedy Leaf Removal on Hypergraphs</td>
<td>94</td>
</tr>
<tr>
<td>9.1</td>
<td>The core of random hypergraphs</td>
<td>97</td>
</tr>
<tr>
<td>9.1.1</td>
<td>core1</td>
<td>98</td>
</tr>
<tr>
<td>9.1.2</td>
<td>core2</td>
<td>100</td>
</tr>
<tr>
<td>9.2</td>
<td>Application</td>
<td>104</td>
</tr>
<tr>
<td>9.3</td>
<td>Discussion</td>
<td>107</td>
</tr>
<tr>
<td>9.A</td>
<td>Critical exponents and critical points</td>
<td>108</td>
</tr>
<tr>
<td>9.A.1</td>
<td>Core1 for uniform hypergraphs - critical point</td>
<td>108</td>
</tr>
<tr>
<td>9.A.2</td>
<td>Core1 - critical exponent</td>
<td>109</td>
</tr>
<tr>
<td>9.A.3</td>
<td>Core2 - critical exponent</td>
<td>111</td>
</tr>
<tr>
<td>9.1.4</td>
<td>Simulations</td>
<td>112</td>
</tr>
<tr>
<td>10</td>
<td>Conclusion</td>
<td>115</td>
</tr>
</tbody>
</table>
Introduction

With the technological advancements during the last decades, we have gathered, documented and mapped enormous data sets – from Instagram photos of cats to the interactions between proteins in a cell. Over the same period as this explosion of data emerged, network science has emerged as a vital new approach to understanding the complexities of networks and complex systems. In 1999 Professor Albert-László Barabási and Professor Réka Albert found that the topology of these networks are not trivial. For example, the distribution of connections per website on the World Wide Web follows a power law with the exponent $\gamma = 2.1 \pm 0.1$ [2]. As a direct consequence, a small fraction of sites contain a much larger number of links than one would expect if the distribution of connections per node followed a typical Poisson distribution. Similar effects were found in many other networks like the protein-protein interaction network [3]. In this thesis we show how network science can also be used to study the cosmic web, and in the later chapters we explore how a generalization of graphs, called a hypergraph, can not only produce a more faithful representation of complex systems, but also make some classic combinatorial problems easier to solve.
CHAPTER 4. INTRODUCTION

4.1 Thesis structure and content

We explore two models for networks embedded in geometric spaces, and we compute their main properties analytically in Chapter 2. In Chapter 3 we present how a network approach can be used to describe the cosmic web. We explore seven network construction algorithms that use various galaxy properties – from their location to their size and relative velocity – to assign a network to galaxy distributions provided by both simulations and observations. We find that a model relying only on spatial proximity offers the best correlations between the physical characteristics of the connected galaxies. We show that the properties of networks generated from the simulations and observations are very similar, reflecting the validity of our theoretical understanding of structure formation in the universe.

In Chapter 4 we explore a hypothetical spreading process at the cosmic level, and we find that the expansion of the universe, combined with the structure of the cosmic web, makes this type of spreading process very hard. The hope that one day mankind will proliferate throughout the universe is deeply embedded in human consciousness, appearing in books, movies and many other types of media. Humanity began to send probes to other planets a little more than a half-century ago [4], and some of these probes are already moving out of our solar system [5]. There are two main limitations to intergalactic travel. The first is the speed of travel, and the second is the distance of travel. Therefore, we can compute an upper limit for the number of reachable galaxies using a civilization that can spread
CHAPTER 4. INTRODUCTION

to a galaxy a distance $l_0$ at a particular velocity $v$. We define a network, taking the nodes as galaxies and links as connections between galaxies along which a civilization can travel. A civilization that starts the spreading process in a galaxy $i$ could only reach a galaxy $j$ if there is a path between them. Because the universe is expanding, $v$ also limits the number of reachable galaxies. If a set of galaxies is moving away from a civilization faster than the civilization is getting closer to them, the agent will never reach those galaxies. Based on data from the Illustrious simulation [6], it is possible to show that the average number of reachable galaxies is much smaller in this spreading process (moving from galaxy to galaxy) than by traveling in a straight line.

Finally, in Chapters 5 and 6 we explore the properties and applications of a generalization of graphs. A hypergraph is a natural generalization of a graph where an edge (often called a hyperedge) can simultaneously connect any number of vertices [7]. The fact that hyperedges can connect more than two vertices facilitates a more faithful representation of many real-world networks. For example, a hypergraph can typically represent collaboration networks, where vertices represent individuals and hyperedges connect individuals who were involved in a specific collaboration (e.g., a scientific paper, a patent, a consulting task, or an art performance) [8, 9]. Hypergraphs are also used to represent cellular networks [10]. Given a set of proteins and a set of protein complexes, the corresponding hypergraph naturally captures information on proteins that occur together in a protein complex. For a biochemical reaction system, the hypergraph representation will indicate which bimoleculars participate in a particular reaction [10, 11]. We offer analytical solutions to
classic percolation problems on hypergraphs with arbitrary vertex degree and hyperedge cardinality distributions. We introduce a generalization of the 2-core for hypergraphs, and we show that it can emerge in either a continuous or a hybrid percolation transition. We also define two different hypergraph cores associated with the hyperedge cover and vertex cover problems. We validate our analytical results with extensive numerical simulations.

We pay special interest to the hypergraph’s cores. The core of a graph – defined as the remainder of the greedy leaf removal (GLR) procedure where leaves (vertices with a degree one) and their neighbors are removed iteratively from the graph – has been related to the conductor-insulator transition \[12\], structural controllability \[13\], and many combinatorial optimization problems \[14\]. In fact, the size of the core is related to a fundamental combinatorial problem: the complexity of the minimum edge cover problem, being either a solvable problem in polynomial time if the relative size of the core is zero or typically a NP-hard problem if the relative size of the core is non-zero \[15, 16\]. We apply our methodology to numerous real networks that can be described as hypergraphs, and we find that both cores were small in the tested networks, making it possible to compute the size of the minimum edge cover set or vertex cover set with a small margin of error in those same networks. For example, for the APS publication network which is composed of nearly all articles published in APS journals from 1893 to 2010, we find that we need between 21.5\% and 21.6\% of all authors to cover all articles.

Throughout the rest of the introduction, we introduce the definitions, nomenclatures and concepts necessary to understand the rest of the thesis.
CHAPTER 4. INTRODUCTION

4.2 Definitions and nomenclature

A complex network, usually described using graphs, is a set nodes (or vertices) connected by links (or edges). We can use this structure to describe complex systems by taking the important elements of a system as nodes – like proteins in a protein-protein interaction...
network and websites on the World Wide Web – and using links to describe interactions, correlations or other important relationship between nodes. Graphs can be written as an adjacency matrix. For the two nodes $i$ and $j$, the adjacency matrix is defined as

$$a_{ij} = \begin{cases} 1 & \text{if there is a link between node } j \text{ and } i, \\ 0 & \text{otherwise}. \end{cases} \quad (4.1)$$

Links in the network don’t have a specific direction in undirected networks: therefore, $a_{ij} = a_{ji}$. In directed networks there are directions for each link (see Figure 4.1).

In an undirected network the degree of a node $i$, $k_i$, is defined as the number of links connected to a node $i$ (given by $k_i = \sum_{j=0}^{N} a_{ij} = \sum_{j=0}^{N} a_{ji}$, where $N$ is the number of nodes in the network). In directed networks there are two types of degrees: the in-degree, $k_i^{\text{in}}$, defined as the number of links incident to a node $i$, $k_i^{\text{in}} = \sum_{j=0}^{N} a_{ij}$; and the out-degree, $k_i^{\text{out}}$, defined as the number of links going from $i$ to other nodes $k_i^{\text{out}} = \sum_{j=0}^{N} a_{ji}$ (see Figure 4.1).

As the number of links in a network increases, clusters of nodes eventually connect to produce connected components. For undirected networks, the connected component is a set of nodes, so that for every two nodes in the set, $i$ and $j$, there is a path connecting $i$ to $j$. For directed networks there are strongly and weakly connected components. A strongly connected component is a set of nodes, and for every two nodes, $i$ and $j$, in the set there is a path between connecting $i$ to $j$. A weakly connected component is a set of nodes, in which every for every pair of nodes, $i$ and $j$, there is a path from $i$ to $j$, ignoring the
CHAPTER 4. INTRODUCTION

direction of the nodes (see Figure 4.1). For undirected networks the strongly and weakly
connected component are the same thing. Let’s consider a network in the thermodynamic
limit $N \to \infty$. For this network the giant component is a component that contains a finite
fraction of nodes. This definition originated in the context of a theoretic model, where
we could test if the growth of a connected component is proportional to the size of the
networks. In real networks we usually measure the largest connected component. While
the giant component is a distinct concept from the largest component, the two are often
used interchangeably because if a giant component exists and it’s unique, it is also the
largest component. While in very heterogeneous networks two giant components might
be possible, it is very unlikely in random networks. If we consider two connected giant
components $G1$ and $G2$, it is very likely that a link between a node in $G1$ and node in $G2$,
exists.

4.3 Classic random network models

Several network models have been proposed that reflect the properties of real networks. In
1959 Paul Erdős and Alfréd Rényi introduced the Erdős-Rényi network that captured the
randomness of real networks [18][19].
4.3.1 Erdős-Rényi model

In an Erdős-Rényi network, links connect two randomly selected nodes. There are two slightly different versions of this model that are equivalent in the thermodynamic limit \( N \to \infty \) \cite{18, 19}. In the version of the model proposed by Alfréd Rény and Paul Erdős the number of links, \( L \), is fixed, and they connect two randomly selected nodes. The average degree is then given by \( \langle k \rangle = \frac{2L}{N} \). Edgar Gilbert proposed an alternative version of random networks in which for every pair of nodes a connection between is established with a probability of \( p \). Gilbert’s approach has an average degree of \( \langle k \rangle = pN \). The degree distribution (probability of finding a node with degree \( k \)) in this model follows a Poisson distribution \cite{18, 19},

\[
P(k) = \frac{k^c e^{-c}}{k!}, \tag{4.2}
\]

where \( c \equiv \langle k \rangle \). Figure 4.2 shows the size of the giant component as a function of the average degree, \( \langle k \rangle \). For \( \langle k \rangle < 1 \), the network is in a disconnected regime (i.e. there is no giant component and the network is composed by small connected components). For \( \langle k \rangle > 1 \), the network is in a connected regime (i.e. there is a giant component in the network).

For a general random undirected network (i.e. networks with a given degree distribution but no correlations), there will be a giant component if

\[
\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} > 1, \tag{4.3}
\]
where $\langle k \rangle$ and $\langle k^2 \rangle$ are the first and second moment distributions, $\langle k \rangle \equiv \sum_{k=0}^{\infty} P(k)k$ and $\langle k^2 \rangle \equiv \sum_{k=0}^{\infty} P(k)k^2$.

**Figure 4.2:** Connectivity regime. Relative size of the giant component, $S_g$, in function of the mean degree $\langle k \rangle$ for an Erdős-Rényi network. The dashed line represents the phase transition from the disconnected to the connected regime.
CHAPTER 4. INTRODUCTION

4.3.2 Barabási-Albert model

The degree distribution in real world networks\cite{2,1} does not follow a Poisson distribution. For example, the World Wide Web is a network composed of nodes (websites) and links (hyperlinks connecting two websites), and the degree distribution follows a power law,

\[ P(k) \propto k^{-\gamma}, \]  

(4.4)

with \( \gamma = 2.1 \pm 0.1 \)\cite{2}. Barabási and his collaborators introduced two concepts which explain this distribution: growing networks and preferential attachment. Let’s start with a small undirected network to which we add nodes and links at each time \( t \). Each new node will be connected to \( n \) nodes in the network. The probability, \( \pi_i \), of a new link being connected to a node \( i \) is proportional to the degree of node \( i \),

\[ \pi_i \propto k_i. \]  

(4.5)

Networks obtained by using this model follow a power law with, \( \gamma = 3 \). A variation of this model, explored by Dorogovtsev and his collaborators, assumes that

\[ \pi_i \propto k_i + A, \]  

(4.6)
where $-n < A < \infty$. In this case $\gamma = 3 + A/n$, and therefore $2 < \gamma < \infty$ \[17\]. This mechanism explains two key features of real-work networks: hubs and the small world effect. Hubs are nodes with a much larger degree than the average degree. They appear in the network because nodes with a large degree have an higher probability of acquiring new nodes than nodes with a smaller degree according to the preferential attachment.

The small world effect, often referred to in popular culture as “six degrees of separation”, describes how in real networks, even if there are many nodes in a network, the distance between them is usually small (around 6). Another key feature of the Barabási-Albert model and variation, which is related to the existence of hubs in the network, is that the second moment of the degree distribution $\langle k^2 \rangle$ diverges when $N \to \infty$ if $2 \leq \gamma \leq 3$. This model explains some very important properties of real-world networks, but it fails to explain other properties like correlations between the degree of connected nodes, community structure and the non-zero cluster coefficient in the thermodynamic limit (see further discussion of this concept in Chapter 2) \[1, 17\].

### 4.4 Network epidemics

Understanding the topology of the underlying network allows us to understand and predict how dynamical processes evolve in a complex system. Let’s explore the impact of network topology in three simple spreading processes: SI, SIS and SIR models.

Let’s consider the SI and SIS model in a network where nodes can only be in two states: susceptible, $S$, or infectious, $I$. In contrast, nodes in the SIR model in a network can be in
CHAPTER 4. INTRODUCTION

three states: susceptible, S, infectious, I, or recovered, R. At any time in three of the models there is a probability, $i_k$, that a susceptible node $k$ will be infected by a neighboring node $j$ with a probability of $\beta$, and in the SIS and SIR models an infectious node also has the probability, $\mu$, of becoming susceptible or recovered, respectively. There is one big difference between the SIS and SIR model. In the SIR model, nodes that have recovered are considered immune and can not be infected again, but in the SIS model those nodes can be infected again. Table 4.1 summarizes the dynamical properties of each model to random networks with no correlations, where $\tau$ is the characteristic time defined as necessary for 100/e percentage of the networks to get infected, $\theta_k$ is the probability that the neighbor of a node with degree $k$ is infected, and $\lambda_c$ is the epidemic threshold rate. If $\lambda \equiv \beta/\mu > \lambda_c$, then the pathogen will die out; and if $\lambda \leq \lambda_c$, the pathogen will continue to spread.

<table>
<thead>
<tr>
<th>Model</th>
<th>Equation for $i_k$</th>
<th>Equation for $\theta_k$</th>
<th>$\tau$</th>
<th>$\lambda_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI</td>
<td>$\frac{di_k}{dt} = \beta (1 - i_k) k \theta_k$</td>
<td>$\sum_{k'} k' \theta_{k'} k' (k')^2 - \sum_k k \theta_k (k)$</td>
<td>$\frac{1}{\beta (k^2 - (k))}$</td>
<td>0</td>
</tr>
<tr>
<td>SIS</td>
<td>$\frac{di_k}{dt} = \beta (1 - i_k) k \theta_k - \mu i_k$</td>
<td>$\sum_{k'} k' \theta_{k'} k' (k')^2 - \sum_k k \theta_k (k)$</td>
<td>$\frac{\beta (k^2 - \mu (k))}{\beta (k^2 - (k))}$</td>
<td>$\frac{\lambda_c}{\langle k \rangle}$</td>
</tr>
<tr>
<td>SIR</td>
<td>$\frac{di_k}{dt} = k \theta_k - \mu i_k$</td>
<td>$\sum_{k'} k' \theta_{k'} k' (k')^2 - \sum_k k \theta_k (k)$</td>
<td>$\frac{\beta (k^2 - \mu (k))}{\beta (k^2 - (k))}$</td>
<td>$\frac{\lambda_c}{\langle k \rangle}$</td>
</tr>
</tbody>
</table>

Table 4.1: Equation for $i_k$, $\tau$ and $\lambda_c$ for the SI SIS and SIR models.

then the pathogen will die out; and if $\lambda \leq \lambda_c$, the pathogen will continue to spread. where $\lambda \equiv \beta/\mu$. The network’s topology plays an important role in this case because for a scale free network where $\langle k^2 \rangle \rightarrow +\infty$, $\lambda_c \rightarrow 0$. Therefore in scale free networks with $2 < \gamma < 3$ (or any network with a diverging second moment), the pathogen spreads very quickly, and there is always some spreading of diseases even if the average degree of the network is very small. There is one intuitive explanation for this effect. Hubs – nodes with
CHAPTER 4. INTRODUCTION

a much larger degree than the usual – help the spread of disease. It is also worth noting that if we set $\mu = 0$ in the SIS or SIR model, the probability of a node getting infected, $i_k$, converges to the probability of a node getting infected in the SI model. We can think of the SI model as an upper bound epidemic model. Every node connected to a component that contains at least one infected node will also end up infected (see Chapter 4 for further details).

4.5 Role of topology in the complexity of combinatorial problems

Network topology does not only impact the dynamical processes. It is also related to the computational complexity of several network measurements and combinatorial problems. Computing certain network properties or solving certain problems is not always trivial or even possible in a reasonable amount of time. For some problems like computing the cluster coefficient [20] or the shortest path [21] of a network, it is possible to find the solution in polynomial time (i.e. if the size of the system doubles, the computational time required to find a solution will increase by a factor of two elevated to some power). For other problems, a small increase in the system’s size will increase the amount time required to find a solution faster than any power, making the problem, in practice, impossible to solve in a reasonable amount of time [22]. Ladner and his co-authors offer a metric to measure the complexity of a problem that is still used today [22]. They divided the problems into
four categories $P$, $NP$, $NP$-complete, and $NP$-hard (see Figure 4.3). Intuitively these four sets of problems can be described as (For a rigorous mathematical definition see [22]):

- **P-problems** can be solved in polynomial times using a deterministic Turing machine.

- **NP-problems** are the set of problems where a proposed solution can be verified in polynomial time using a deterministic Turing machine.

- **NP-hard problems** are at least as hard as the hardest $NP$ problem using a deterministic Turing machine.

- **NP-complete problems** belong to the intersection between $NP$ and $NP$-hard problems (i.e. the set of problems that are at least as hard as the hardest $NP$ problem but a proposed solution can still be verified in polynomial time using a deterministic Turing machine).

While it is generally assumed that $NP \neq P$, this has never been proven [23]. Figure 4.3 shows a diagram of these sets under the two assumptions $P = NP$ and $P \neq NP$. Remarkably, the complexity of certain measurements can change depending on the properties of the network. Let us consider the vertex-cover problem of an undirected network, $G$. The minimum vertex-cover problem consists of finding the smallest set of vertices, $S$, so that at least one edge is incident to a node in the set $S$. Figure 4.4 (a) shows a network where blue nodes can cover all the edges; it is also the smallest set needed to cover all the edges. Finding the smallest set of vertices that covers all edges is generally a NP-hard problem [24, 25], but for some nodes we can use simple rules to determine whether or not...
they belong to a solution of the minimum edge-cover problem, $S$. Let’s look at nodes $v_1$, $v_2$ and edge $e_1$ in Figure 4.4 (a). $e_1$ can only be covered by nodes $v_1$ or $v_2$, but because $v_2$ covers edges $e_1$, $e_2$ and $e_3$, $v_2$ is always as good as or a better option to cover $e_1$ using $v_2$ than $v_1$. Therefore, there is at least one solution for the minimum edge-cover problem, where $v_2$ belongs to the set $S$, $v_1$ doesn’t belong to the set $S$, and edges $e_1$, $e_2$ and $e_3$ are covered by $v_2$. We can remove those nodes and links from the networks, and now we only need to solve the minimum vertex-cover problem in the reminder of network $G$ without nodes $v_1$ and $v_2$, or edges $e_1$, $e_2$ and $e_3$. This process, often called the greedy leaf removal process [26], can be repeated until no more nodes can be removed from the network, and
CHAPTER 4. INTRODUCTION

Figure 4.4: Core of a network. Considering the solution \( S \), for the minimum edge-cover problem obtained by the greedy leaf removal process: blue nodes belong to \( S \); black nodes do not; red nodes are part of the core. We cannot determine whether or not red nodes belong this minimum edge-cover solution. Figure (a) is an example of a network with \( S_c = 0 \); therefore, the minimum-edge cover problem can be solved in polynomial time. Figure (b) is an example of a network with \( S_c \neq 0 \); therefore, the minimum-edge cover problem can be solved in polynomial time using this algorithm. Figures (c) and (d) are examples of two distinct solutions for the minimum edge-cover problem which the same network.

the remainder network is called the core of the network [26]. It can be summarize as:

- Consider an edge \( e_1 \), which connects to two nodes \( v_1 \) and \( v_2 \). If the degree of \( v_1(v_2) \) is equal to one but the degree of \( v_2(v_1) \) is larger than one, \( v_1, v_2 \) and all edges connected to \( v_2(v_1) \) can be removed from the network. Furthermore, \( v_2(v_1) \) belongs to the minimum edge-cover set, \( S \), while \( v_1(v_2) \) does not.
CHAPTER 4. INTRODUCTION

• Consider an edge $e_1$ which is connected to two nodes $v_1$ and $v_2$. If the degree of $v_1$ and $v_2$ are equal to one, one of them belongs to the minimum edge-cover set, $S$, and the other will not. Because $v_1$ and $v_2$ are connected to exactly the same edge, we can randomly select one of them.

• We repeat this process until no more nodes can be removed from the network.

From Figure 4.4 it is easy to see the size of the core is zero. The minimum vertex-cover problem can be solved in polynomial time. If the core is not zero, there is no known general way to solve this problem in polynomial time (However, for some particular cases like bipartite networks, there are polynomial time solutions [24]). Figure 4.4 (a) is an example of a network where the minimum edge-cover problem can be solved using this method, and Figure 4.4 (b) is an example where it can’t. It is important to notice that if the size of the core is zero, we obtain a solution for the minimum edge cover problem, but not necessarily the only solution. Figure 4.4 (c) and (d) shows two possible solutions for the minimum edge cover problem in the same network. The core in contrast is unique. For an Erdős-Rényi network this problem can sometimes be solved in polynomial time, depending on the average degree $c$. Figure 4.5 shows the size of the core as a function of $c$. For $c \leq e$ the relative size of the core is zero. For $c > e$ the relative size of the core is finite (where $e$ is the Euler’s number $e \approx 2.71828$). This corresponds to a P to NP-hard transition in the complexity of the problem. This transition is very similar to the second order phase
transition found in physical systems, described as,

$$S_c \sim (c - e)^\gamma$$  \hspace{1cm} (4.7)

for $c - e \to 0^+$ and $\gamma = 1/2$.

Yang-Yiu and his collaborators found that the size of the core is always zero if the degree distribution of a network is a pure power law, making the vertex-cover problem always solvable in those networks [26]. We can intuitively understand why this is the case. If the degree distribution of a network follows a power law, there is large fraction of nodes with a degree of one, making it very likely that a hub will be connected to at least one

![Network core in function of ⟨k⟩.](image)

Figure 4.5: **Network core in function of ⟨k⟩.** Relative size of the network core $S_c$, in function of the mean degree $\langle k \rangle$ for an Erdős-Rényi network. The dashed line represents the phase transition.
CHAPTER 4. INTRODUCTION

node with degree one and if the hubs of network are removed, then the entire network will collapse into small connected components. Yang-Yiu and his collaborators also found that in a directed network the core can emerge as a first order phase transition,

\[
S_c = \begin{cases} 
  a + B(c - c^*)^\gamma & \text{if } c - c^* \to 0^+ \\
  0 & \text{if } c < c^*
\end{cases} 
\] (4.8)

where \( c^* \) is the phase transition critical mean degree, \( a > 0, B > 0 \) and \( \gamma < 1 \). The vertex-cover problem is related to other combinatorial problems (e.g. maximum matching \[14\]). Therefore, it follows that transitions in computational complexity are present in other combinatorial problems. Phase transitions in the computational complexity of a problem were also independently found in 2-Sat and related problems \[27\] (See Chapter 7 for further discussion).

4.6 Large scale cosmic simulations

While network topology plays an important role in what problems can we solve and it impacts the spread of diseases, no network representation is complete and can capture every detail about a system. The key purpose of network science is finding the representation of a complex system that captures its essential properties but is not so overwhelmingly detailed that it loses its effectiveness. This approach produced has been extremely useful in biology \[28,29,30\] and the social sciences where problems are extremely complex \[31\].
CHAPTER 4. INTRODUCTION

Cosmologists are now facing a similar problem. During the last half decade, advancements in computation have made simulating the evolution of the universe much more accurate [32]. We can run different models of the universe and compare them to the real universe, offering a new way to find the model that best describes the evolution of the universe. Notable, Crain and his collaborators produced a large scale simulation in 2009 [33], and the recent Illustris simulation [32 34] was the first large scale simulation of the universe that took into account both dark and baryonic matter, producing a much more realistic universe. While this simulation opened a new way to test models and assumptions concerning the formation of the universe, it also created a new set of problems related to comparing simulations with observation data. Because the observed data is incomplete and because the definitions of the objects in the simulation do not always match observed objects in the universe this task is not trivial, and we need a new set of tools to systematically complete this process. In Chapter 3 we propose to apply network science to study the comic web, which could produce a set of tools to address these problems in the future.
Spatial Networks

Networks embedded in metric spaces, usually called spatial networks or geometric graphs, are present in our daily life, and embedding networks in a geometric graphs often impacts the properties of those same networks [35, 36, 37]. One example is the power grid, where power cables connect houses and factories to power plants [38]. These networks tend to be mostly planar, meaning that they can be projected into two dimensional Euclidean space in such a way that there is little or no crossing between links. The reason for this is simple. Power cables crossing close to each other is undesirable, and using a three dimensional construction to avoid crossing is not ideal. Similarly, in the brain, neurons are connected by axons, forming a network embedded in a three dimensional space. In this case regions of the brain that are far away have a lesser chance of being connected than regions nearby since long axons are more expensive in terms of energy [39]. While for some networks it is easy to see how the spatial information constrains the links between nodes, this effect is less obvious in other networks. For example, in a friendship network it is much more likely for people that live in the same city or nearby cities to befriend each other than
people that live very far away [35]. The way spatial information constrains links in this network is not trivial. These networks are usually called spatial networks, and how spatial information affects these networks has been of interest to the scientific community for the last decade [35]. Some spatial network models create networks with properties similar to real world networks. This suggests that spatial information might have a direct or indirect role in most real world networks. In this chapter we explore two models of spatial networks in order to better understand how spatial networks differ from other networks.

5.1 Random geometric graph

In this section we will explore a simple well-studied model, the geometric graph [40, 41, 42]. In this model, nodes are are embedded in $d$-dimensional euclidean space. Two nodes are linked if the distance between them is smaller than a certain value. Let’s consider this model when applied to a random distribution of nodes, embedded in a 2-dimensional space with a length of one for simplicity. The average number of nodes at distance $r$ from a node 0 is given by [43]

$$\rho(r) \, dr = 2\pi nr \, dr,$$  \hspace{1cm} (5.1)

where $n$ is the density of nodes. If we consider that two nodes are connected when the distance between them is smaller than $R$, then the average degree of a node is given by [42]

$$\langle k \rangle = \int_0^R \rho(r)\,dr = n\pi R^2.$$  \hspace{1cm} (5.2)
CHAPTER 5. SPATIAL NETWORKS

We can compute the second moment of the degree distribution by considering node density at distance \( r_2 \) of a node 0, given the fact that there is a node at distance \( r_1 \),

\[
\rho(r_2|r_1) \, dr = n\pi r_2. \tag{5.3}
\]

Note that the existence of a node at distance \( r_1 \) from a node 0 doesn’t change the density of points at distance \( r_2 \). Using this result one can compute the average number of triplets around node 0 (see Figure 5.1) as,

\[
\langle k(k-1) \rangle = \frac{\langle k^2 \rangle - \langle k \rangle}{2} = \frac{1}{2} \int_0^R dr_2 \int_0^R dr_1 \rho(r_2|r_1)\rho_2(r_2) = \langle k \rangle^2. \tag{5.4}
\]

That holds

\[
\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle. \tag{5.5}
\]

We can compute the distribution’s higher moments from

\[
\langle \prod_{i=0}^{n-1} (k-i) \rangle = \langle \frac{k!}{(k-n)!} \rangle = \prod_{i=1}^n \int_0^R d \, r_i \, \rho(r_i|r_1, \ldots, r_{i-1}), \tag{5.6}
\]

where \( \rho(r_i|r_1, \ldots, r_{i-1}) \) is the probability of finding a node at distance \( r_i \) from another node, assuming that there are already nodes at distances \( r_1, \ldots, r_{i-1} \) from a certain node.

As before, the existence of nodes at distance \( r_1, \ldots, r_{i-1} \) from node 0 doesn’t change the
Figure 5.1: Number of triplets centered at node 0. This value can be computed in two ways: by counting all the possible combinations, in this case $(3 \times 2)/2 = 3$; or by integrating the probability of finding one node at distance $r_2$ and the other at distance $r_1$ over all the possible values of $r_1$ and $r_2$.

density of nodes at distance $r_i$. Therefore, the density of nodes at distance $r_i$ remains the same,

$$\rho(r_i|r_1, \ldots, r_{i-1}) = 2 \pi r_i.$$  (5.7)

Combining equations 5.6 and 5.7, we obtain

$$\langle \frac{k!}{(k-n)!} \rangle = \langle k \rangle^n.$$  (5.8)

These moments correspond to the moments of a Poisson distribution. Thus, it follows that
CHAPTER 5. SPATIAL NETWORKS

the degree distribution of this network is given by

\[ P(k) = \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!}, \quad (5.9) \]

where \( P(k) \) is the probability of finding a node with degree \( k \). While there are easier ways to derive this result (see [42]), the advantage of this approach will be become clear in Section 5.2. This result can be generalized to any euclidean space with an arbitrary dimension \( d \) by considering the average degree of a node as [42]

\[ \langle k \rangle = \frac{n\sqrt{\pi R^d}}{\tau \left( \frac{d+2}{2} \right)}. \quad (5.10) \]

While the degree distribution in this model and the Erdős-Rényi model are the same, other properties are not. For example, the giant component emerges at \( \langle k \rangle = 1 \) for the Erdős-Rényi model, but it emerges at \( \langle k \rangle = 4.52, 2.74 \) or 2.06 for \( d = 2, 3, 4 \) for random geometric graphs, respectively [42]. Another important difference is the value clustering coefficient in the thermodynamic limit. The clustering coefficient is defined as the average probability that two nodes connected to a node \( j \) are also connected each other. The clustering coefficient of an Erdős-Rényi network converges to zero for infinitely large networks while the networks generated by this model have a finite coefficient for finite dimensions. This is of special interest to us. While most models like the Erdős-Rényi model and Barabási-Albert model produce networks with a cluster coefficient equal to zero in the thermodynamic limit [1, 44], most real world networks have a non-zero cluster coefficient.
even in the thermodynamic limit [45]. Scholars believe that spatial information is one of the mechanisms that produces networks with a non-zero cluster coefficient and that graph models embedded in metric space can mimic the properties of real world networks. For example, Dmitri Krioukov and his collaborators have proposed a model using geometric graphs in hyperbolic 2-dimensional space that can produce networks with a power law degree distribution and a non-zero cluster coefficient. Those networks are similar to what is found in real world networks [46]. It is also important to notice that for infinite dimensions, \( d \to \infty \), all the properties of random geometric graphs converge to properties of an Erdős-Rényi network [42].

5.2 random k-nearest neighbor graph

A slightly more complex and less studied geometric graph is the \( k \)-nearest neighbors graph. In the \( k \)-neighbor graph, there are connections between nodes and their \( k \)-nearest neighbors. This model and its variations are widely used for optimization algorithms, pattern recognition, machine learning, data mining and non-linear regression among many other applications [47, 48, 49, 50]. Despite its popularity, much less is known about this model than the model described in Section 5.1, and even its basic properties, like the emergence of a giant component, are still being studied [51]. In order to gain a better understand of this model, I applied it to a random distribution of points in \( 2D \), focusing mainly on \( k = 1 \). While this is the simplest form of this model, it exemplifies the main problems when deriving analytic expressions. Namely, that the position of each node constrains the links of
nearby nodes. As discussed before, the probability of finding \( m \) nodes at a distance \( r \) from a node follows a Poisson distribution,

\[
P_{GG}(m) = \frac{\langle m \rangle^m e^{-\langle m \rangle}}{m!}, \tag{5.11}
\]

where \( \langle m \rangle \equiv n\pi r^2 \) and \( n \) are the density of nodes. Now assume that there is a node 0 and a node \( i \) at a position \( \mathbf{x} \) in relation to node 0. If a node \( i \) is connected to node 0, there can be at most \( k - 1 \) nodes at distance less than \( r \equiv |\mathbf{x}| \) from node 0 (see Figure 5.3). Therefore, we can derive the probability that there is a link between node 1 and 0 as

\[
\tilde{\rho}_k(\mathbf{x}) = n e^{-n\pi x^2} \sum_{s=0}^{K-1} \frac{(n\pi x^2)^s}{s!}.
\tag{5.12}
\]

If we define the vector \( \mathbf{x} \) in polar coordinates as \( \mathbf{x} \equiv r(\cos(\theta), \sin(\theta)) \), we can obtain the average number of neighbors at a distance \( r \) by integrating equation 5.12 over all the possible values of \( \theta \) (see Figure 5.2):

\[
\rho(r) \equiv \int_0^{2\pi} d\theta \tilde{\rho}_k(r \cos(\theta), r \sin(\theta)) = 2n\pi r e^{-n\pi r^2} \sum_{s=0}^{K-1} \frac{(n\pi r^2)^s}{s!}.
\tag{5.13}
\]

The average in-degree of a node is given by

\[
\int_S d\mathbf{x} \; \tilde{\rho}_k(\mathbf{x}) = \int_0^{\infty} dr \; \rho_k(r) = k.
\tag{5.14}
\]
Chapter 5. Spatial Networks

Figure 5.2: 1-nearest neighbor graph, distance distribution. (a) Represents the probability of finding a link between two nodes at distance \( r \), \( \rho_1(r) \), for \( \langle k \rangle = 1 \), and \( n = 4.76 \times 10^{-11}, 2.00 \times 10^{-11} \) and \( 5.93 \times 10^{-12} \). (b) Represents the probability of finding a link between two nodes at distance \( r \), \( \rho_k(r) \), for \( n = 1 \), \( \langle k \rangle = 1 \), \( \langle k \rangle = 2 \) and \( \langle k \rangle = 3 \). The solid black line present the theoretical curves. The blue, green and red lines are obtained from simulations in networks with 20,000 nodes.

This result is to be expected since the average in-degree is equal to the average out-degree, in this case \( k \) in any graph.

It is also important to notice that we can set \( n = 1 \) without losing generality. Setting \( n = 1 \) is equivalent to replacing \( r \to r/\sqrt{n} \), and it is easy to check that

\[
\rho_k(r) dr = \rho_k \left( \frac{r}{\sqrt{n}} \right) d \left( \frac{r}{\sqrt{n}} \right).
\] (5.15)
5.3 Moments of the degree distribution for k=1

While all nodes have the same out-degree, that is not true for the in-degree. In order to compute the variance of the in-degree distribution, let’s assume the existence of a third node \( k \). Nodes 1 and 2 will be at positions \( x_1 \) and \( x_2 \), respectively, in relation to node 0. Furthermore, for simplicity, we assume that the distance between node 2 and 0 is bigger than the distance between node 1 and 0, \( |x_2| > |x_1| \). The existence of a link between node 1 and node 0 changes the probability of having a link between node 2 and node 0 in two ways:

- The distance between node 2 and node 0 needs to be smaller than the distance between node 2 and 1, \( |x_2 - x_1| > |x_2| \) (see Figure 5.3).

- If we assume \textit{a priori} that node 1 is connected to node 0, we assume \textit{a priori} that there are no nodes in a radius \( |x_1| \) from \( x_1 \). Likewise, in order for node 2 to be connected to node 0 there can be no nodes in a radius \( |x_2| \) from \( x_0 \), but we already assume that there are no nodes in a radius \( |x_1| \) from \( x_1 \). Therefore, we should not take into consideration the intersection between those two areas when computing the probability of having no other node closer to 2 than 0.

Considering this, the probability of having node 2 connected to node 0, assuming that node 1 is connected to node 0, is given by

\[
\tilde{\rho}_1 (x_2|x_1) = e^{-(\pi|x_2|^2-A(x_2|x_1))} \Theta(|x_1-x_2| - |x_2|) \Theta(|x_2| - |x_1|).
\]  \hspace{1cm} (5.16)
CHAPTER 5. SPATIAL NETWORKS

\( \Theta(x) \) is the Heaviside step function, where \( \Theta(x) = 0 \) if \( x < 0 \) and \( \Theta(x) = 1 \) if \( x > 1 \).  
\( A(x_2|x_1) \) is the area of the intersection of a circle or radius \( |x_2| \) around \( x_2 \), with a circle of radius \( |x_1| \) around \( x_1 \):

\[
A(x_2|x_1) \equiv |x_1|^2 \cos^{-1} \left( \frac{d^2 + |x_1|^2 - |x_2|^2}{2d|x_1|} \right) + |x_2|^2 \cos^{-1} \left( \frac{d^2 + |x_2|^2 - |x_1|^2}{2d|x_2|} \right)
+ \sqrt{(-d + |x_1| - |x_2|)(-d + |x_2| - |x_1|)(-d + |x_2| + |x_1|)(d + |x_2| + |x_1|)},
\]

(5.17)

where we define,

\[
d \equiv |x_1 - x_2|.
\]

(5.18)

Using this result, we can compute the average number of triples per node as

\[
\frac{\langle k_{in}(k_{in} - 1) \rangle}{2} = \frac{\langle k_{in}^2 \rangle - \langle k_{in} \rangle}{2} = \int_S d\mathbf{x}_2 \int_S d\mathbf{x}_1 \bar{\rho}_1(\mathbf{x}_2|\mathbf{x}_1)\bar{\rho}_1(\mathbf{r}_2),
\]

(5.19)

where \( \langle k_{in} \rangle \) and \( \langle k_{in}^2 \rangle \) are the first and second moment of the in-degree distribution, respectively. Since \( \langle k_{in} \rangle^2 = \langle k_{in} \rangle = 1 \), we obtain

\[
\langle k_{in}^2 \rangle = 1.6322.
\]

(5.20)

We confirmed this result using 10 randomly generated networks with \( 10^4 \) nodes, and we
CHAPTER 5. SPATIAL NETWORKS

Figure 5.3: **Condition for the connections between two and three nodes for** \( \langle k \rangle = 1 \):
(a) In order for a node 1 to be connected to a node 0 at distance \( r \), there can be no node in a region of radius \( r \) around node 1 (indicated by the brown region around node 1). (b) There are two conditions in order for a second node 2 to be connected to node 0, assuming that node 1 is already connected to 0. First, node 2 cannot be in the brown region around node 1. Second, there can be no nodes in the blue region around node 2, excluding the intersection between the blue and brown regions. The condition that node 1 is connected to 0 already implies *a priori* that there are no nodes in the brown region.

obtained \( \langle k_{in}^2 \rangle = 0.6311 \pm 0.0047 \) in agreement with our calculations. The variance of this distribution is \( \langle k_{in}^2 \rangle - \langle k_{in} \rangle^2 = 0.6332 \), smaller than the variance of a Poisson distribution with the same average degree (in this case 1). Our approach can be easily be generalized to other dimensions, and for 3D, we obtained an in-degree distribution with a variance equal to \( \langle k_{in}^2 \rangle \approx 1.709 \).
CHAPTER 5. SPATIAL NETWORKS

We can use this approach to compute the higher moments of the distribution. In general,

\[
\langle \frac{k_{\text{in}}!}{(k_{\text{in}} - n)!} \rangle = n! \prod_{i=1}^{n} \int_S d\mathbf{x}_i \bar{\rho}_1(\mathbf{x}_i|\mathbf{x}_1, \ldots, \mathbf{x}_{i-1})
\]

(5.21)

where \( \bar{\rho}_1(\mathbf{x}_i|\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}) \) is the probability of having a link between node \( i \) at position \( x_i \), and node 0. Given the fact that there are already \( i - 1 \) nodes connected to node 0 from positions \( x_1 \) to \( x_{i-1} \) and the distance between node \( i \) and node 0, \( |x_i| \), is bigger than the distance between node \( j \) and node 0, \( |x_j| \), for \( j = 1, \ldots, i - 1 \)

\[
\bar{\rho}_1(\mathbf{x}_i|\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}) = e^{-\left(\pi|x_i|^2 - A(\mathbf{x}_i|\mathbf{x}_1, \ldots, \mathbf{x}_{i-1})\right)} \times \theta(|x_i| - |x_{i-1}|) \prod_{j=1}^{i-1} \theta(|x_j - x_i| - |x_i|).
\]

(5.22)

(5.23)

Similarly, \( A(\mathbf{x}_i|\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}) \) is the area of the intersection between a circle with a radius \( |x_i| \) around \( x_i \) for any sphere with a radius \( |x_j| \) around nodes, for \( j = 1, \ldots, i - 1 \).

\[
A(\mathbf{x}_i|\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}) \equiv \int_S d\mathbf{x} \theta(|x_i| - |\mathbf{x} - x_i|) \times \prod_{j=1}^{i-1} \theta(|\mathbf{x} - x_i| - |x_i|).
\]

(5.24)

(5.25)

Typical intersections can be solved using a Monte Carlo approach, but because we have the exponential of \( A(\mathbf{x}_i|\mathbf{x}_1) \), integrating all possible values of \( x_1, x_2, \ldots, x_i \) would force us to use that approach for every value of \( x_1, x_2, \ldots, x_i \), making the integration very slow. We can avoid this problem by combining a Taylor expansion and a finite number of integrals.
CHAPTER 5. SPATIAL NETWORKS

using the Monte Carlo approach. For example, take the third moment:

\[
\langle (k_{in} - 2)(k_{in} - 1)k_{in} \rangle = \int_S d\mathbf{x}_1 \int_S d\mathbf{x}_2 \int_S d\mathbf{x}_3 \tilde{\rho}_1(\mathbf{x}_3|\mathbf{x}_2, \mathbf{x}_1)\tilde{\rho}_1(\mathbf{x}_2|\mathbf{x}_1)\tilde{\rho}_1(\mathbf{x}_1) \quad (5.26)
\]

While we don’t have an explicit expression for \( \tilde{\rho}_1(\mathbf{x}_3|\mathbf{x}_2, \mathbf{x}_1) \), we can expand it as

\[
\tilde{\rho}_1(\mathbf{x}_3|\mathbf{x}_2, \mathbf{x}_1) = e^{\pi|x_3|^2} \Theta(|x_3| - |x_2|)\Theta(|x_3 - x_1| - |x_3|)\Theta(|x_3 - x_2| - |x_3|) \quad (5.27)
\]

\[
\lim_{T \to \infty} \sum_{t=0}^{T} \left( \int_S d\mathbf{x} \Theta(|x_3| - |x - x_3|)\Theta(|x - x_2| - |x_2|)\Theta(|x - x_1| - |x_1|) \right)^t / t!.
\]

By combining this equation with equation 5.26 we can obtain a sum of simple multidimensional integrals that can easily be solved using the Monte Carlo approach. This same approach can compute any of the moments. We confirm our results using simulations: Figure 5.4 shows that by increasing \( T \) the third moment \( \langle k_{in}^3 \rangle \), which was obtained from equations 5.26 and 5.28, converges to the value obtained in simulations. For the fourth moment we calculated \( \langle k_{in}^4 \rangle = 6.6770 \), which agrees with the simulations where \( \langle k_{in}^4 \rangle = 6.6 \pm 0.1 \) ( obtained from 10 random generated network with \( 10^4 \) nodes). We also found that for \( m > 4 \),

\[
\prod_{i=1}^{m} \int_S d\mathbf{x}_i \tilde{\rho}_1(\mathbf{x}_i|\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}) = 0, \quad (5.29)
\]

meaning that the maximum in-degree of a node is equal to 4 for \( k = 1 \). This result is confirmed in the simulations, where we don’t find any nodes with a degree larger than 4.
Figure 5.4: 1-nearest neighbor graph, third moment of the in-degree distribution. \( \langle k^3_{in} \rangle \) in function of \( T \). The black line represent the average value of \( \langle k^3_{in} \rangle \), and the green region represents the region within 1 standard deviation, obtained from 10 random generated network with \( 10^4 \) nodes. The red line shows the value of \( \langle k^3_{in} \rangle \) computed from equations 5.26 and 5.28 in function of \( T \).

Since we know all the moments of the in-degree distribution, we can compute it using

\[
\sum_{k=0}^{4} k^np_{in}(k) = \langle k^n \rangle \tag{5.30}
\]

or

\[
\begin{bmatrix}
1 & 1 & 1 & 1 \\
0 & 1 & 2 & 3 & 4 \\
0 & 1 & 4 & 9 & 16 \\
0 & 1 & 8 & 27 & 64 \\
0 & 1 & 16 & 81 & 256
\end{bmatrix}
\begin{bmatrix}
P_{in}(0) \\
P_{in}(1) \\
P_{in}(2) \\
P_{in}(3) \\
P_{in}(4)
\end{bmatrix}
= 
\begin{bmatrix}
1 \\
\langle k_{in} \rangle \\
\langle k^2_{in} \rangle \\
\langle k^3_{in} \rangle \\
\langle k^4_{in} \rangle
\end{bmatrix} \quad . \tag{5.31}
\]
CHAPTER 5. SPATIAL NETWORKS

Figure 5.5: In-degree distribution for the 1-nearest neighbors model. In-degree distribution, \( P(k_{in}) \), of the k-nearest neighbors model applied to a two dimensional random distribution of points, computed using eq.5.31 and from 10 random generated network with 10^4 nodes. We also plot a Poisson distribution for comparison.

This system of equations can be easily solved. Figure 5.5 shows that the theoretical in-degree distribution and the simulated in-degree distribution agree.

The cluster coefficient for \( k = 1 \) is equal to zero since it is impossible to create a closed triangle between three nodes with one link per node. Let’s consider the 3 nodes in Figure 5.6 where, for simplicity and without losing generality, we assume that distance between node 2 and node 0 is bigger than the distance between node 1 and node 0. In this case node 0 is necessarily connected to node 1. In order to have a closed triangle, a link must exist from node 1 to 2. Therefore, the distance between node 1 and 2 need to be smaller than the distance between node 1 and node 0, marked as the red region in Figure 5.6. To complete the triangle, node 2 needs to be closer to node 0 than node 1, but by design,
every position in the red region is closer to node 1 than node 0. Therefore, closed triangles are impossible in this model. For a similar reason, there can be no giant component for \( k = 1 \). Let’s look to node 3 in Figure 5.6. Node 3 represents a neighbor of node 2 that is not node 1 or node 0. This type of node can only be found in the region marked as green, which is smaller than the brown region. Similarly, the region where it is possible to find node 4, a neighbor of node 3 that is not node 0, 1 or 2, will be smaller than the green region. This is a strong indication that there is no giant component in this model for \( k = 1 \). This observation, was proven by Teng and Yao [52] and Bagchi and Bansal [53], who showed that in 2D the giant weakly connected component emerges for \( 2 < \langle k \rangle < 188 \). This result was later improved by Balister and his collaborators who proved that \( \langle k \rangle < 188 \) [54]. Using simulations, H. Häggström and his collaborators found that the weakly connected giant component emerges at \( \langle k \rangle = 3 \) for \( d = 2 \) and \( \langle k \rangle = 2 \) for \( d > 2 \) [55]. This model is a very clear example of how geometry can constrain the properties of a network. In this case, the position of each node constrains the links of nearby nodes, resulting in an in-degree distribution with smaller variance in an Erdős-Rényi network. Figure 5.7 shows that, by increasing the dimensions of the space, the variance tends to be 1, equal to the variance of the degree distribution in an Erdős-Rényi network and a random geometric graph [42]. This result is to be expected since the geometric constraints present in this model become less relevant for large dimensions.
Figure 5.6: Closed triangles and a giant connected component in the KNN model with $\langle k \rangle = 1$: Let us consider 3 nodes, so that node 0 is connected to node 1 and node 1 is connected to node 2. The brown region represents the only possible region where node 1 is connected to node 2. The dashed arrow represents the last links necessary to close the triangle. Nevertheless, node 2 will always connect to node 1 first, and there are no close triangles in this model. Node 3 represents a neighbor of node 2 that is not node 1 or node 0, and node 4 represents a neighbor of node 3 that is not node 0, 1 or 2. The green and blue regions represent, the regions where we can find such nodes, respectively. Note that these regions get smaller and smaller as we add more nodes.
Figure 5.7: **Variance of the KNN model with \( \langle k \rangle = 1 \) for different dimension.** Variance \( \langle k_{in}^2 \rangle - \langle k_{in} \rangle^2 \) of the In-degree distribution of the KNN model for \( \langle k \rangle = 1 \), in a \( D \)-euclidean space in function of \( D \). The dash line shows the variance of the degree distribution for a Erdős-Rényi network. These values were computed using 10 independent simulations of networks with 20,000 nodes, for each value of \( D \).
The Network Behind the Cosmic Web

The cosmic web, imagining the large-scale structure of the universe as a network, is deeply embedded in cosmology and public consciousness [32, 34, 56, 39, 57]. Yet, it remains little more than a metaphor, typically used to described dark matter’s ability to agglomerate the galaxies in a web-like fashion. Thanks the increasingly precise simulations of the universe’s evolution [32, 58], numerous halo finder algorithms [59, 60] exploit the network-like binding of galaxies [61], but, very little is known about the graph theoretical characteristics of the resulting cosmic web. Our goal is to test and explore various meaningful definitions of the cosmic web and characterize the generated networks with the tools of network science. In particularly, we explore which network definition offers the best description of the observed correlations between the physical characteristics of connected galaxies. The resulting network-based framework, tested in both simulations and observational data, offers a new tool with which we can investigate the topological properties of the large-scale structure distribution of the universe.
6.1 Simulated and observational galaxy samples

We chose samples from a state-of-the-art hydrodynamic cosmological simulation, the Illustris simulation [32, 34], and observational data from the Sloan Digital Sky Survey (SDSS) [62, 63]. In line with common practice, we assume that subhalos in the simulation correspond to galaxies in the observational data [64], representing the nodes of the cosmic web.

The Illustris simulation starts approximately 2 million years (Myr) after the Big Bang and computes the evolution of the universe until the present epoch in a volume of $(75\text{Mpc}/h)^3$, taking into account both dark and baryonic matter with a resolution of $6.26 \times 10^6 M_{\odot}$ and $1.26 \times 10^6 M_{\odot}$. It is the first cosmological simulation that finds a mix of blue and red galaxies with a metallicity in agreement with observations. From the Illustris snapshots # 60, 80, 100 and 135 (at redshifts $z = 3.01$, $z = 1.21$, $z = 0.578$ and $z = 0$, respectively), we chose all simulated galaxies with stellar masses $M^* \geq 10^9 M_{\odot}$.

The SDSS is a very wide-field imaging and spectroscopic survey of the sky that used the 2.5-m wide-angle telescope located at Apache Point Observatory in Sunspot, New Mexico. From the SDSS Data Release 7, we chose galaxies with stellar masses $M^* \geq 10^9 M_{\odot}$ and right ascension $100 < RA < 270$ and declination $-7 < DEC < 70$. Table 6.1 shows the number of galaxies, number density, average distance between all galaxies, volume of each galaxy, and shape of the region of the space selected for the two data sets. While the SDSS has many incompletenesses due to fiber collisions and bright stars among other issues, for the purpose of this initial study we ignored these selection effects since they typically only
affect sampling in the densest clusters. At a redshift of 0.03, 1 arcsecond corresponds to a scale of 0.6 kpc. This means that the typical inter-galaxy distance is \( \sim 1 \) Mpc, or 0.5 deg, which is much larger than the scale of any of the incompletenesses.

<table>
<thead>
<tr>
<th>Number of galaxies</th>
<th>Volume ( (\text{Mpc}/h)^3 )</th>
<th>Redshift</th>
<th>Galaxy number density ( (\text{Mpc}/h)^3 )</th>
<th>Space geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDSS</td>
<td>19,280</td>
<td>( 7.73 \times 10^4 )</td>
<td>( z &lt; 0.03 )</td>
<td>( 2.49 \times 10^{-2} )</td>
</tr>
<tr>
<td>Illustris snap060</td>
<td>4,995</td>
<td>( 4.22 \times 10^4 )</td>
<td>3.01</td>
<td>( 1.18 \times 10^{-2} )</td>
</tr>
<tr>
<td>Illustris snap080</td>
<td>16,250</td>
<td>( 4.22 \times 10^4 )</td>
<td>1.21</td>
<td>( 4.22 \times 10^{-2} )</td>
</tr>
<tr>
<td>Illustris snap100</td>
<td>21,086</td>
<td>( 4.22 \times 10^4 )</td>
<td>0.578</td>
<td>( 5.00 \times 10^{-2} )</td>
</tr>
<tr>
<td>Illustris snap135</td>
<td>24,305</td>
<td>( 4.22 \times 10^4 )</td>
<td>0</td>
<td>( 5.76 \times 10^{-2} )</td>
</tr>
</tbody>
</table>

Table 6.1: Basic information about the data from Illustris, SDSS. The number of galaxies is roughly the same in the data from the Illustris and SDSS selection, but because the Illustris overestimates the stellar mass function for low mass galaxies, the number density of galaxies in the SDSS is smaller than in the Illustris simulation.

Figure 8.A.3(a) shows the galaxy stellar mass function, which is defined as the number of galaxies with stellar mass, \( M_\star \), per volume, for the data sets used in the main text. A known feature of the Illustris simulation [65] is that the resulting \( z = 0 \) galaxy stellar mass function is in excess of the observed mass function at the low mass end (i.e., for \( M_\star < 10^9 \) \( M_{\odot} \)); hence, we only consider samples with a stellar mass \( M_\star \geq 10^9 \) \( M_{\odot} \). Subhalos in Illustris containing galaxies with a stellar mass of \( M_\star \geq 10^9 \) \( M_{\odot} \) are well resolved. They are significantly above the mass resolution of the simulation. Figure 8.A.3(b) reveals that the mass resolution mainly affects subhalos with a stellar mass smaller than \( 10^6 \) \( M_{\odot} \). We obtain a sample of 4,995, 16,250, 21,086 and 24,205 galaxies from the Illustris simulation at redshifts \( z = 3.01, z = 1.21, z = 0.578 \) and \( z = 0 \), respectively.
Figure 6.1: (a) Represents galaxy stellar mass functions for observed and simulated galaxies at $z = 0$. (b) Stellar mass versus dark matter mass for the simulated subhalos at the redshifts $0$, $0.578$, $1.21$ and $3.01$. The color axis represents the fraction of points in each pixel for a grid with $100 \times 100$ pixels. We only consider galaxies with a stellar mass $M_* \geq 10^9 M_{\odot}$, marked as a vertical dashed line in (a) and a horizontal dashed line in (b).
6.2 Models

There are multiple ways to build networks from the available subhalo/galaxy catalogs, allowing us to define seven distinct models for the construction of the cosmic web (M1-M7). The simplest, M1, links two nodes with an undirected link if the distance between them is smaller than a predefined length, \( l \). M2(3) represents the directed versions of M1, drawing a directed link \( j \rightarrow i \) (\( i \rightarrow j \)) between \( i \) and the closest \( \langle k \rangle \) nodes. Consequently, while a node can have arbitrary degree in M1, the in(out) degrees are fixed in M2(3). In M4(5) a directed link \( j \rightarrow i \) (\( i \rightarrow j \)) is drawn between \( i \) and \( j \) if the distance between the two nodes is smaller than \( a \cdot \frac{R_1^{1/2}}{2} \), where \( a \) is a free parameter and \( R_1^{1/2} \) is the half-mass radius \([66]\), defined as the radius of the sphere centered on the subhalo containing half of the mass of the subhalo. Model 6(7) is an extension of M4(5), but is computed in phase space, where a directed link \( j \rightarrow i \) (\( i \rightarrow j \)) is drawn between \( i \) and \( j \) if the sum of the square of the distance and relative speed between two nodes when divided by the peak circular velocity radius and peak circular velocity, is smaller than \( a^2 \). The peak circular velocity is the maximum value of the circular velocity profile of the subhalo and the peak circular velocity radius is the radius that maximizes the circular velocity \([67]\). Taken together, M1-3 require only data about the halo/galaxy positions; M4 and M5 require information about the galaxy masses; additionally, M6 and M7 require galaxy peculiar velocities, attempting to only link together galaxies that are gravitationally bound to each other and ignoring those that come close but eventually separate. The adjacency matrix and information required of each model is given
Figure 6.2: **Building networks from galaxy data.** The circles represent the linking lengths for models M1, M3 and M4. (a) In M1 all galaxies within distance \( l \) are connected by an undirected link. (c) In M3 a galaxy is connected to the closest galaxy with a directed link; therefore, the linking length depends on the position of the closest galaxy. (e) In M4 the linking length scales with each galaxy size: \( l = a R_{1/2}^i \). (b),(d) and (f) are visualizations of the cosmic web for redshift 0 produced by each respective model for \( \langle k \rangle = 40 \). For simplicity, the direction of the links is not present in the visualization. For the interactive visualization, see [http://kimalbrecht.com/ccnr/04-networkuniverse/17-network-interface](http://kimalbrecht.com/ccnr/04-networkuniverse/17-network-interface). Models M2, 5, 6, 7 are generated from the three models shown above. In M2 the directions of the M3 links are inverted; in M5 the direction of the M4 links are inverted. M6(7) are similar to M4(5) but they are computed in phase space.

in Table 6.2
Table 6.2: The summary of features for each network model; the required information from galaxies, algorithm to define the adjacency matrix, and the type of generated network. \( a_{ij} = 1 \) means a link between galaxy \( j \) and \( i \), otherwise \( a_{ij} = 0 \). We denote the half-mass radius, peak circular velocity radius, velocity and peak circular velocity of galaxy \( i \) by \( r_{1/2}^i, r_{\text{max}}^i, v_i \) and \( v_{\text{max}}^i \), respectively. \( d_{ij} \) is the distance between galaxy \( i \) and \( j \).

Table 6.2 summarizes the properties of each network model used to characterize the galaxy point distribution.

### 6.3 Pairwise correlations

The distinct network representations of the cosmic web, offered by the models introduced above, raise the question: Which of these representations is the most meaningful? In general, networks are only meaningful if the links have functional roles, linking either interacting nodes or nodes with similar characteristics. For example, the links of a social network tend to connect individuals with similar social-economic characteristics (homophily), and in cellular networks connected proteins tend to have related biological
roles. The fact that the color of a satellite galaxy is correlated with the mass of the host galaxy \[68, 69, 70, 71, 72, 73\] indicates that such correlations between nearby galaxies are meaningful. Therefore, we explore the degree to which the above network representations of the cosmic web add links between galaxies/subhalos of similar physical characteristics. For this we analyze 71 parameters that characterize each subhalo, ranging from their peculiar velocity to star formation rate, allowing us to identify the network representation that offers the best correlation between them. Since we are working from a cosmological simulation, some of the correlations may be meaningful only in the sense that they characterize the underlying properties (or assumptions) of the model. Nevertheless, our analyses provides an unbiased way of probing the spatial network without any \textit{a priori} biases.

For a given model M and subhalo property \( c_i \), we compute the average value of \( c_i \) over all nodes connected to \( i \),

\[
\tilde{c}_i \equiv \frac{\sum_j a_{ij} c_j}{k_i},
\]

(6.1)

where \( k_i \) is the degree of node \( i \) and \( a_{ij} \) is the adjacency matrix. We use the Pearson coefficient to measure correlations between the connected nodes,

\[
R \equiv \frac{\sum_i (\tilde{c}_i - \langle \tilde{c} \rangle)(c_i - \langle c \rangle)}{\sqrt{\sum_i (\tilde{c}_i - \langle \tilde{c} \rangle)^2 \sum_i (c_i - \langle c \rangle)^2}},
\]

(6.2)

where \( \langle c \rangle \) and \( \langle \tilde{c} \rangle \) are the average of \( c_i \) and \( \tilde{c}_i \) over all nodes. Since the scale over which correlations persist is unknown, we construct networks with different average degrees, \( \langle k \rangle \).

We identify the strongest negative or positive correlation between the galaxy’s proper-
ties, \( R_{\text{max}} \),

\[
\begin{aligned}
R_{\text{max}} &= \max_{0<(k)<40} R(\langle k \rangle) \text{ if } \max_{0<(k)<40} R(\langle k \rangle) \geq \left| \min_{0<(k)<40} R(\langle k \rangle) \right| \\
R_{\text{max}} &= \min_{0<(k)<40} R(\langle k \rangle) \text{ if } \max_{0<(k)<40} R(\langle k \rangle) < \left| \min_{0<(k)<40} R(\langle k \rangle) \right|
\end{aligned}
\]

(6.3)
Figure 6.3: **Correlation.** $R_{\text{max}}$ for models 1-7 applied to the simulation and all properties tested. The properties discussed in the text are marked in bold.

We find four properties that consistently display correlations between the connected nodes: peculiar velocity, stellar metallicity, specific star formation rate, and color in the B-V band (Figure 6.4). We also calculated the correlation function for networks obtained under node and link randomization (see Section 6.4). The lack of significant correlations in these randomized networks indicates that Figure 6.4 captures physical meaningful correlations between connected galaxies. While all models captures physical meaningful correlations to some degree, M3 captures the best correlations for the peculiar speed, specific
star formation rate, and rest-frame B-V color. It only fails to maximize the correlation between stellar metallicities, in which M6 excels. Note that some of these correlations, like metallicity, may result from the assumptions made by the simulation, while others, like peculiar velocity, likely reflect physical correlations in the real universe. However, this exercise demonstrates that it is indeed possible to uncover underlying properties of the network without prior knowledge. Overall, Figure 6.4 indicates that model M3 captures best the correlations between the properties of connected galaxies. Its superiority over the more data demanding models M4-M7 suggests that spatial proximity, despite its simplicity, remains the most powerful organizing principle of the cosmic web.

6.4 Randomized networks

By definition, the Pearson coefficient of any property is close to zero in a randomized version of the network (Fig 6.5). In this section we confirm this result for the four explored properties. We consider two different types of randomization:

**Node properties randomization** - the links of each node are maintained, but the properties are randomized;

**Degree-preserving edge randomization** - the properties of each node are maintained, but the links are randomized in a way that the degree of each node is preserved.
Figure 6.4: Correlations between connected galaxies. The Pearson coefficient $R$ capturing the correlation between connected nodes as a function of the mean degree $\langle k \rangle$ for all algorithms, redshift 0 and for various galaxy properties. The giant strongly connected component emerges at $\langle k \rangle = 4$ for M3, shown as a dashed vertical line.
Figure 6.5: Correlations between connected galaxies in randomized networks. The Pearson coefficient $R$ captures the correlation between connected nodes as a function of the mean degree $\langle k \rangle$ for M3 and various galaxy properties in randomized networks. We compute the mean value and standard deviation using 10 independent randomizations.
CHAPTER 6. THE NETWORK BEHIND THE COSMIC WEB

6.5 M3 applied to real data

Given the ability of M3 to best capture correlations between the subhalo characteristics, in the remainder of the paper we analyze the networks predicted by this model.

The overall integrity of a network is well characterized by the size of its largest connected component \([1]\). For a directed network the strongly connected component is the largest subset of nodes such that for all pairs \(i\) and \(j\) in the subset there is a directed path from \(i\) to \(j\). Figure 6.6 illustrates that the giant strongly connected component emerges at \(\langle k \rangle = 4\) for all redshifts in M3. At the practical level, this implies that M3 can be applied at different redshifts and numbers of nodes without the need to adjust the model parameters. At a more fundamental level, it means that the critical mean degree of the giant strongly connected component is universal, being rooted in the intrinsic proprieties of the galaxy distribution.

To further validate M3, we compare the structure of the cosmic web obtained in the simulations with observational data from the Sloan Digital Sky Survey (SDSS) \([62, 63]\), that provides information about the position and the properties of the galaxies in the visible sky \([65]\). We study the section of the sky with redshift \(z < 0.03\), right ascension \(100 < RA < 270\) and declination \(-7 < DEC < 70\). As a reference, we also study a randomized version of the data by distributing the galaxies randomly in space (Appendix G). To confirm that the galaxy population in the simulation and in the observations are comparable in the same redshift range, we measured the galaxy mass function, defined as the number
of galaxies with stellar mass $M_\ast$ per volume. Since the density of galaxies with stellar mass under $10^9 M_{\odot}$ is much higher in the simulation than in the observational data, a known limitation of the simulation [65], we only consider galaxies with stellar mass $M_\ast > 10^9 M_{\odot}$.

Figure 6.6(a) documents an excellent agreement between the in-degree distribution (fraction of nodes with a given in-degree), $P(k_{\text{in}})$, for the observational and the simulated M3 networks. Both distributions deviate from the random distribution, indicating that the observed $P(k_{\text{in}})$ reflects the non-trivial galaxy distribution in both the simulations and in the observations. We show analytically in supplementary material D, that networks constructed by the M3 model for a random galaxy distribution, the variance of the in-
degree distribution is 0.709 for $\langle k \rangle = 1$. The fact that the variance is smaller than 1, which is the value expected for a Poisson distribution, implies that the degree distribution is narrower than that expected for an Erdős-Rényi network. Therefore, hubs are definitely absent; this is quite distinct from what is found in biological and social networks, where hubs are prevalent. We also obtain excellent agreement between the simulation and observation based networks for the average clustering coefficient, capturing the fraction of triangles in the network, and assortativity (Figure 6.7(c) and (d)). In both cases the simulation and observation-based values agree with each other, both deviating from the random expectation. The giant strongly connected component emerges at $\langle k \rangle = 4$ for both the simulation and observational M3 networks, while it is at $\langle k \rangle = 3$ for the random M3 networks (Figure 6.7(b)). The results indicate that M3 offers an accurate description of the cosmic web, capturing consistently its network characteristics, both in the simulation and in the observational data.
Figure 6.7: **Network characteristics of the cosmic web.** (a) shows the in-degree distribution for mean degrees $\langle k \rangle = 1, 4$ and 20. The dashed lines represents the Poisson distribution. (b) shows the size of the largest strongly connected component, $S_g$, as a function of the mean degree, $\langle k \rangle$. (c) represents the average clustering coefficient as a function of the average mean degree $\langle k \rangle$. (d) displays the assortativity coefficient, $r$, as a function of $\langle k \rangle$. Each panel show M3 data for the random, simulated (redshift 0) and the observational networks.
6.6 Conclusion

In summary, here we used the tools of network science to characterize the large structure of the universe both in simulations and observational data. While we can define numerous network construction algorithms, we find that the simple model M3, which relies on spatial proximity only, captures the best correlations between the physical characteristics of nearby galaxies. The results are distinct from the random case, which assumes random galaxy localizations, indicating that the obtained structure of the cosmic web is intricately tied to the underlying structure of the universe. It is particularly encouraging that the network characteristics of the cosmic web, from the degree distribution to the clustering and degree correlations, show remarkable agreement between simulations and observations. In many ways, our results represent only the first step towards a network-based understanding of the universe. Yet, they provide guidance for the nature of the data needed for a systematic exploration of the underlying network, offering a framework on which one could build various applications, from halo finders to exploring the fundamental characteristics of the cosmic web.
Spreading Across the Universe

The hope that one day mankind will proliferate through the universe is deeply embedded in human consciousness, manifesting itself in books, movies and many other types of media.

We started to send probes to other planets little more than half-century ago [4], and some of these probes are already moving out of our solar system [5]. Some hypothesize that in the near future we will be able to send probes capable of self-repairing and self-replicating [6], maximizing our chances of reaching other solar systems or, in the distant future, other galaxies. If we desire to send probes or humans throughout the universe, we need to strategize an appropriate path. In this work we explore the most likely scenario to perform such exploration. We find that in most cases, only a small fraction of the universe is reachable. This result might be at the core of the famous “Fermi paradox”, which relates to the apparent contradiction between the likely existence of other civilizations in the universe but the fact that we cannot find any evidence of them.
7.1 Model

There are two main limitations to intergalactic travel. The first is the speed of travel since all evidence suggests that traveling at velocities equal to or greater than the speed of light is physically impossible. Second, spending colossal amounts of time in virtually empty regions of the universe, with limited resources and without the chance of exploiting them, makes travel extremely difficult. It stands to reason that intelligent and non-intelligent civilization would have a limited distance that they can travel before being destroyed. Due to this problems, intelligent civilizations or life forms would be discouraged from trying to travel longer distances than they must. The most likely scenario is that any intelligent life form would prefer to maximize their chances of spreading locally to galaxies at a distance they know they can reach safely, rather than risking spending longer distances in empty space than necessary. For lack of a better work, I’m going to use the term civilization to anything that spreads from one galaxy to the other, namely, an actually civilization, some non-intelligent life or probes capable of self-repairing and self-replicating. We can model this process by using some civilization that can spread to a galaxy to a distance $l$, at velocity $v$. Since agents can only travel to a galaxies at distance $l$ limits the number of galaxy they can reach, we can use percolation theory to quantify the fraction of reachable galaxies. Let’s define a network where nodes represent galaxies, and links represent a civilization that can travel between two galaxies (i.e, the distance between them is smaller than the distance $l$ measured a certain time, $t_0$). It is important to mention that a galaxy $i$ can only
reach a galaxy $j$ if there is a path between them. If the universe was static, the size of the connected component would be the number of reachable galaxies for any civilization that is contained in that component. Figure 7.1 shows the size of the giant component in function of the linking length. Nevertheless, the universe is not static, and due to the expansion of the universe, galaxies are moving apart. Therefore the linking length that creates a path between two galaxies at time $t_0$ is not the same length that will create a path between those same two galaxies after some time $t$. Let’s now consider a linking length $l_0$ and the associated set of pairs of nodes $A$, composed by all pairs of nodes such that there is a path between them in the network as measured at some time $t$. Let’s also consider the linking length $l$, defined as the smallest linking length that makes it possible for a civilization to travel between any of the pair of nodes in $A$ at velocity $v$, (taking into account that galaxies are moving further way from each other). Using Hubble’s law and, one can show that $l$ and $l_0$ are related to each.

### 7.2 Traveling time

We know from Hubble’s law that changes in distance between galaxy $i$ and $j$, $D_{ij}(t)$ can be approximated by

$$\frac{d}{dt} D_{ij}(t) = H_0 D_{ij}(t). \quad (7.1)$$
Figure 7.1: Giant component as a function of $l_0$. Relative size of the giant component as a function of $l_0$. The orange dashed vertical line marks the point in which the curves $\gamma(l_0)$ and $A(l_0)$ change behavior.

If a civilization, $s$, travels from galaxy $i$ to $j$ at velocity $v$, the change in the distance between the civilization $s$ and galaxy $j$, $D_{sj}(t)$, is given by

$$\frac{d}{dt} D_{sj}(t) = D_{sj}(t) - v.$$  \hspace{1cm} (7.2)
CHAPTER 7. SPREADING ACROSS THE UNIVERSE

For now, let’s consider that $H_0$ is constant in time. We can derive that the time required to travel from galaxy $i$ to $j$, $T_{ij}$ is,

$$T_{ij} \equiv \int_{D_{ij}(t_1)}^{0} \frac{dD_{sj}}{H_0D_{st} - v}. \quad (7.3)$$

We obtain

$$e^{-H_0T_{ij}} = 1 - \frac{D_{ij}(t_1)H_0}{v}. \quad (7.4)$$

Let’s now assume that, after arriving at galaxy $j$ (at time $t_2 \equiv t_1 + T_{ij}$) the civilization then travels to galaxy $k$. The traveling time, from galaxy $j$ to $k$ is

$$e^{-H_0T_{kj}} = 1 - \frac{D_{kj}(t_2)H_0}{v}. \quad (7.5)$$

The distance between galaxy $k$ and $j$ at time $t_2$ is related to the distance between galaxy $k$ and $j$ at time $t_1$ is

$$D_{jk}(t_2) = D_{jk}(t_1)e^{H_0T_{ij}}. \quad (7.6)$$

By combining equations 8.20, 8.19, and 8.23,

$$e^{-H_0T_{ik}} = 1 - \frac{D_{ik}(t_1)H_0}{v} \quad (7.7)$$

where $T_{ik}$ is the time to travel from $i$ to $k$, $T_{ik} \equiv T_{ij} + T_{jk}$ and $D_{ik} \equiv D_{ij} + D_{jk}$. One can generalize this result to any number of trips. Because this formula is equivalent to equation
CHAPTER 7. SPREADING ACROSS THE UNIVERSE

we conclude that we can apply this formula not only for trajectories in a straight line, but also for the motion inside of a network.

\[ e^{-H_o T_{ij}} = 1 - \frac{D_{ij} H_o}{v}, \quad (7.8) \]

where \( T \) is the traveling time and \( D_{ij} \) is the initial distance between the two galaxies. From this equation, it is easy to see that there is a maximum distance that a civilization can travel (since the LHS of the equation is always positive). Depending on the velocity of the civilization, we obtain that the maximum distance an civilization can travel is

\[ D_{\text{max}} = \frac{v}{H_o}. \quad (7.9) \]

This equation is valid if the civilization travels in a straight line, or if it goes from galaxy to galaxy. The theoretical reachable universe is considered to be the region of the universe that one can reach by traveling in a straight line, but traveling in a such manner would force the civilization to travel through completely empty regions of the universe, and as we argued before this is not a plausible scenario. If we consider the far more likely scenario where a civilization only travels to a galaxy at a certain distance, forming the network described above, then the network distance is much more important than the physical distance.

Let’s now consider that the rate of expansion of the universe changes with time, i.e. \( H_o = H_o(t) \). In this case,

\[ D_{\text{max}} = vK, \quad (7.10) \]
CHAPTER 7. SPREADING ACROSS THE UNIVERSE

where

\[ K = \int_{0}^{+\infty} e^{-\int_{0}^{t} H_0(t') dt'} dt. \] (7.11)

If we take into account the changes on the rate expansion of the universe over time, the distance \( D_{\text{max}} \) will remain proportional to \( v \). However, the constant of proportionality \( K \), will depend on an integral over time involving \( H_0(t) \). For a universe where \( H_0(t) \) is constant in time, we recover the previous result, \( K = 1/H_0 \), and for a universe that is not expanding, \( H_0(t) = 0 \), the constant of proportionality between \( D_{\text{max}} \) and \( v \) diverges: \( K \to +\infty \).

7.3 Linking length in a expanding box

Let’s consider a geometric graph in a expanding box that follow the Hubble’s law, i.e,

\[
\begin{align*}
    a_{ij}(t_0, l_0) &= 1 & \text{if } D_{ij}(t_0) < l \\
    a_{ij}(t_0, l_0) &= 0 & \text{if } D_{ij}(t_0) \geq l,
\end{align*}
\] (7.12)

where \( D_{ij}(t_0) \) is the geometric distance between node \( i \) and \( j \), that will evolve with time as

\[ D_{ij}(t_0) = D_{ij}(0)e^{H_0 t_0}. \] (7.13)

Because the distance between nodes is always increasing, some links will disappear with time. Therefore, even if there is a patch between node \( i \) and \( j \) at time \( t_0 \), there is not
CHAPTER 7. SPREADING ACROSS THE UNIVERSE

necessarily a path at time \( t_0 + T \). Let’s now consider two nodes \( i \) and \( j \), so that there is a path \( P \) between \( i \) and \( j \) for linking length \( l_0 \) and time \( t_0 \). It takes a certain time, \( T \), to travel the shortest path between \( P \) at velocity \( v \). If we consider the linking length

\[
l' = l_0 e^{H_0 T},
\]  

we can conclude that it is always possible to travel from \( i \) to \( j \). This is true, because if \( a_{ij}(t_0, l_0) = 1 \), it follows that \( a_{ij}(t, l') = 1 \) for \( t_0 < t < t_0 + T \). Therefore, all links \((k, m)\), belonging to the path between \( i \) and \( j \) for linking length \( l_0 \) and time \( t_0 \), are also present during the time necessary to travel the path \( P \) if we consider the linking length \( l' \).

Nevertheless, this is only an upper bound. It is possible that, by increasing the value of \( l \), new paths, will appear and some of those paths will reduce the length of the shortest path between \( i \) and \( j \) of the network, and the time require to travel between those two nodes.

Let’s consider the network in figure\[8.A.3\] and two linking length \( l_0 = l_1 \), and \( l_0 = l_2 \equiv 2 \ l_1 \). If \( l_0 = l_1 \), there there is no link between nodes \( i \) and \( j \), and the diameter of the network is \( D_{1\text{diameter}} = 7 \ l_0 \). If \( l_0 = l_2 \), there is a link between \( i \) and \( i \) and \( j \), and the diameter of the network is \( D_{2\text{diameter}} = 4 \ l_0 \). If we take into account the expansion of the universe, from eq\[7.14\] we obtain two upper bounds that guarantee that is possible to to travel to any node in the component, given by

\[
l'_1 = \frac{l_1}{1 - \frac{7 \ l_1 \ H_0}{v}}
\]  

(7.15)
and

\[ l_2' = \frac{2l_1}{1 - \frac{4l_1}{v}} H_0. \tag{7.16} \]

Where \( l_1'(l_2') \) is computed using equation \[7.4\] for \( l_o = l_1(l_2) \). It is easy to check this if \( 7 l_1 H_0 < v < 10 l_1 H_0 \), then \( l_2' < l_1' \). This happens because a longer linking length allows for new paths that can lower the diameter of the network. Therefore, the smallest value of \( l \) that makes possible for a civilization to travel between any two points in any connected component in a network, \( a_{ij}(t_0, l_0) \), is between

\[ l_o \lesssim l \lesssim l_o e^{H_o T_{\text{max}}}, \tag{7.17} \]

where \( H_0 \) is the Hubble’s constant \[74\] and \( T_{\text{max}} \) is the time necessary to travel the longest distance in the network computed at time \( t_0 \). It is important to note that if \( T_{\text{max}} \ll \frac{1}{H_0} \), then this interval is narrow. \( 1/H_0 \) is the age of the universe, therefore, this only becomes a problem for larges amount of time.

\( l_o \) is a more convenient quantity than \( l \) because doesn’t depend on \( T_{\text{max}} \), and we can always compute the actual value of \( l \) after we know the exact path. If the traveling time is much smaller than the age of the universe, \( T_{\text{max}} \ll \frac{1}{H_0} \), it is a small correction. The time that it will take for a civilization to travel from a galaxy \( i \) to a galaxy \( j \) can also be computed using the Hubble’s law.
7.4 Results

We apply this approach to the Illustris simulation. Figure 7.3(a) shows the average fraction of reachable galaxy as a function of $D_{\text{max}}$ (and velocity) when moving in a straight line or through the network. Figure 7.3(b) shows the ratio between those two numbers. We find that the ratio between the average fraction of reachable galaxies by moving in the network and in a straight line roughly scales as

\[
\frac{\langle n^{\text{Net}} \rangle}{\langle n^{\text{Phy}} \rangle} \sim D_{\text{max}}^{\alpha}.
\]  

(7.18)
CHAPTER 7. SPREADING ACROSS THE UNIVERSE

Figure 7.3: Number of reachable galaxies. (a) Average number of reachable galaxies as a function of $D_{\text{max}}$ for motion in the network with different values of $l_o$, when moving in a straight line $l_o = \infty$. (b) Fraction between the average number of reachable galaxies by a civilization moving in the network and in a straight line for different values of $l_o$.

and Figure 7.4 shows $\gamma \equiv 3 + \alpha$ as a function of $l_o$. Since $\gamma < 3$ and $\alpha < 0$, the fraction of galaxies a civilization can reach when moving in the network and in a straight line tends to zero as $D_{\text{max}}$ increases. Therefore a spreading process at the cosmic level is most likely localized in a small region of the universe. It is important to notice that the change in behavior in $\gamma$ not being consistent with the birth of the giant component and the existence of critical behavior away from the critical point, is reminiscent of a Griffith phase transition [75].

7.5 Discussion

According to our model, any spreading process in the universe will be localized in a small region of the universe, as a consequence the idea of humans or any other civilization pro-
liferating through the universe is very unlikely. While our model is very simplistic, any improvement on our model, like introducing the probability that our civilization will die or not reach its destination, will only make these problems more severe.

Figure 7.4: **Scaling as a function of the linking length.** Parameter $\gamma(l_0)$ as a function of $l_0$. The dashed vertical line marks the change in behavior.
Percolations on Hypergraphs

In the previous chapters we have shown how geometric graphs can be used to address a very important problem and an hypothetical spreading process at the cosmic level. In this chapter, we will go beyond graphs and explore hypergraphs, the natural generalization of a graph such that an edge can simultaneously connect any number of vertices [7]. As in graphs, the degree of a vertex in a hypergraph is the number of hyperedges that connect to it. The number of vertices connected by a hyperedge is called the cardinality of that hyperedge. If all hyperedges have the same cardinality $K$, the hypergraph is said to be uniform or $K$-uniform. Note that a graph is just a 2-uniform hypergraph. The fact that hyperedges can connect more than two vertices facilitates a more precise representation of many real-world networks. For example, collaboration network can be typically represented by a hypergraph, where vertices represent individuals and hyperedges connect individuals who were involved in a specific collaboration (e.g., a scientific paper, a patent, a consulting task, or an art performance [8, 9]). Many cellular networks can also be represented by hypergraphs [10]. For example, given a set of proteins and a set of protein complexes,
of the corresponding hypergraph naturally captures the information on proteins that occurred together in a protein complex. For a biochemical reaction system, the hypergraph representation will indicate which bimoleculars participate in a particular reaction [10, 11].

Despite the ubiquity of hypergraphs in social and biological systems, the fundamental structural properties of hypergraphs are not fully understood. Most previous work focused on uniform hypergraphs, ignoring the fact that hyperedges could have a wide range of cardinalities. In this work, we systematically study the percolation transitions on hypergraphs with arbitrary vertex degrees and hyperedge cardinality distributions. We are particularly interested in the emergence of a giant component, the $K$-core. Those special subgraphs have been extensively studied in the graph case and play very important roles in many network properties [76, 77]. The giant component of a graph is a connected component that contains a constant fraction of the entire graph’s vertices, which is relevant to structural robustness and the resilience of networks [44, 78]. The $K$-core of a graph is obtained by recursively removing vertices with degrees less than $K$. (Note that the edges incident to removed vertices are naturally removed during this process. The $K$-core has been used to identify influential spreaders in complex networks [79].

We can naturally extend the definition of giant component to the hypergraph case. Yet, to obtain the $K$-core in a hypergraph, we have to specify how to remove hyperedges containing vertices of degree less than $K$. To achieve that, we introduce the $(K, S)$-core, defined as the largest fraction of the hypergraph where each hyperedge contains at least $S$ nodes in the subset and each vertex belongs to at least $K$ hyperedges in the subset. The
(K, S)-core is obtained by recursively removing vertices with degrees less than K and hyperedges with cardinality less than S.

In this work we offer analytical solutions to the classical percolation problems on hypergraphs with arbitrary vertex degree and hyperedge cardinality distributions. We confirm all our results using numerical simulations.

Figure 8.1: Percolations. Example of the different percolations studied in this section. (a) shows the original hypergraph. (b), (c), (d).
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

8.1 Giant component

A giant component of a hypergraph is a connected component that contains a constant fraction of the entire hypergraph’s vertices. In the mean field picture, we can derive a set of self-consistent equations to calculate the relative size of the giant component, using the generating function formalism \[80\]. Let \( \mu \) represent the probability that a randomly selected vertex from a randomly chosen hyperedge is not connected via other hyperedges with the giant component. Dually, let \( \psi \) represent the probability that a randomly chosen hyperedge connecting to a randomly chosen vertex is not connected via other vertices with the giant component. Then we have

\[
\mu = \sum_{k=1}^{\infty} Q_n(k) \psi^{k-1} \tag{8.1}
\]

\[
\psi = \sum_{r=1}^{\infty} Q_h(r) \mu^{r-1}. \tag{8.2}
\]

Here \( Q_n(k) \equiv kP_n(k)/c \) is the excess degree distribution of vertices (i.e., the degree distribution for the vertices in a randomly chosen hyperedge). \( P_n(k) \) is the vertex degree distribution, and \( c = c_1 \) is the mean degree of the vertices. In general, we define \( c_m \equiv \sum_{k=0}^{\infty} k^m P_n(k) \). \( Q_h(r) \equiv rP_h(r)/d \) is the excess cardinality distribution of hyperedges (i.e., the cardinality distribution for the hyperedges connected to a randomly chosen vertex). \( P_h(r) \) is the hyperedge cardinality distribution, and \( d = d_1 \) is the mean cardinality of the hyperedges. In general we define \( d_m \equiv \sum_{r=0}^{\infty} r^m P_h(r) \).
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

The relative size of the giant component is then given by

\[ s_g = 1 - \sum_{k=0}^{\infty} P_n(k)\psi^k. \]  

(8.3)

Figure 8.2 shows the analytical result of \( s_g \) as a function of the mean degree \( c \) for hypergraphs with Poisson vertex degree distribution and different hyperedge cardinality distributions. Clearly the giant component in hypergraphs emerges as a continuous phase transition with scaling behavior

\[ s_g \sim (c - c^*)^\eta \]  

(8.4)

for \( c - c^* \to 0^+ \), where \( c^* \) is the critical value of mean degree (i.e., the percolation threshold) and \( \eta \) is the critical exponent associated with the critical singularity.

The condition for the percolation transition can be determined by differentiating both sides of Eq. (8.1) over \( \mu \) and then evaluating at \( \mu = 1 \), yielding

\[ \frac{d_s - d}{d} \frac{c_s - c}{c} > 1 \]  

(8.5)

(see section 8.A for details.) Note that a similar relation has been found for uniform hypergraphs [81]. In the graph case (\( d = 2 \) for all edges) we recover the classical result \( \frac{c_s - c}{c} > 1 \) [44, 82].

The critical exponent \( \eta \) can be calculated by expanding equations (8.3) in powers of \( (c - c^*) \) around the critical point \( c^* \). For hypergraphs with bounded moments of cardinality
distribution and degrees distribution, we obtain the same exponent $\eta = 1$ as in the graph case [44, 83].

Figure 8.2: **Giant component as a function of** $c$. The relative size of the giant component, $s_g$, as a function of the mean degree $c$ for hypergraphs with Poisson vertex degree distributions. (a) $d$-uniform hypergraphs with $d = 2, 3, 4$; (b) Hypergraphs with Poisson hyperedge cardinality distribution and mean cardinality $d = 2, 3, 4$. 
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

8.2 \((K, S)\)-core

The \((K, S)\)-core of a hypergraph is obtained by recursively removing vertices with degrees less than \(K\) and hyperedges with cardinality less than \(S\). A hyperedge with cardinality \(r\) is removable if at least \(r - S + 1\) vertices connected to it are also removable and a vertex with degree \(k\) is removable if at least \(k - K + 1\) hyperedges connected to it are also removable.

One can remove a vertex or a hyperedge from the hypergraph and see what is the probability of removing a neighboring hyperedge or vertex, respectively. This allows us to derive a set of self-consistent equations:

\[
\alpha = \sum_{k=1}^{\infty} Q_n(k) \sum_{l=k+1-S}^{k-1} \binom{k-1}{l} \delta^l (1 - \delta)^{k-1-l}, \tag{8.6}
\]

\[
\delta = \sum_{r=1}^{\infty} Q_h(r) \sum_{l=r+1-K}^{r-1} \binom{r-1}{l} \alpha^l (1 - \alpha)^{r-1-l}. \tag{8.7}
\]

where \(\alpha\) and \(\delta\) are, respectively, the probability that a vertex or a hyperedge is removable.

From now on we will focus on the case of \(K = 2\). Then Equation (8.6) reduces to

\[
\alpha = \sum_{k=1}^{\infty} Q_n(k) \delta^{k-1}. \tag{8.8}
\]
8.2.1 \( K = 2 \) and \( S = S_{\text{max}} \)

The \((K, S)\)-core defined with \( K = 2 \) and \( S = S_{\text{max}} \equiv \max(r, 2) \) is obtained by recursively removing all vertices with degree one as well as the hyperedges containing them and all hyperedges with cardinality smaller than two. Hyperedges with cardinality one or zero do not connect any nodes and thus have no meaning in which cores are concerned. In this case the threshold \( S \) depends on the cardinality of the hyperedges. If one of the vertices of any hyperedge is removed the hyperedge is also removed. (Note that the \((2, S_{\text{max}})\)-core has been defined in literature \cite{84} simply as 2-core, and a discontinuous 2-core percolation is found in \( d \)-uniform hypergraphs with \( d > 2 \). In this case, equation (8.7) reduces to

\[
1 - \delta = \sum_{r=2}^{\infty} Q_h(r) (1 - \alpha)^{r-1}.
\] (8.9)

The relative size of the \((2, \max(r, 2))\)-core is given by the probability that a randomly chosen vertex is connected to at least two non-removable hyperedges:

\[
s_{2c} = \sum_{k=2}^{\infty} P_h(k) \sum_{l=2}^{k} \binom{k}{l} (1 - \delta)^{\delta^{k-l}}.
\] (8.10)

Figure 8.4 shows the analytical result of \( s_{2c} \) as a function of the mean degree \( c \) for hypergraphs with Poisson vertex degree distribution and different hyperedge cardinality distributions. We find that, depending on the mean hyperedge cardinality \( d \), the \( 2, S_{\text{max}} \)-core
emerges as either a continuous or a hybrid phase transition with scaling behavior

\[ s_{2c} - s_{2c}^* \sim (c - c^*)^\zeta \]  \hspace{1cm} (8.11)

for \( c - c^* \to 0^+ \), where \( c^* \) is the 2-core percolation threshold and \( \zeta \) is the critical exponent. \( s_{2c}^* \) is the two-core relative size at the critical point: \( s_{2c}^* = 0 \) for continuous phase transitions and non-zero for hybrid phase transitions. The percolation threshold \( c^* \) can be calculated by differentiating both sides of equation (8.8) over \( \alpha \) and then evaluating the critical point. It yields

\[ 1 = \sum_{k=1}^{\infty} \sum_{r=2}^{\infty} Q_n^*(k)Q_h(r)(k-1)(r-1)\delta^{k-2} \delta^{r-2}(1-\alpha^*)^r. \]  \hspace{1cm} (8.12)

The critical point for the 2, \( \max(r, 2) \)-core percolation is given by

\[ Q_h(2) \frac{c_2^* - c^*}{c^*} = 1. \]  \hspace{1cm} (8.13)

The phase transition is continuous if

\[ 2Q_h(3)(c_2^* - c^*) - Q_h(2)^2(c_3^* - 3c_2^* + 2c^*) < 0. \]  \hspace{1cm} (8.14)

(See section 8.A for details.). For \( d \)-uniform hypergraphs the 2, \( \max(r, 2) \)-core percolation is (i) continuous with critical exponent \( \zeta = 2 \) if \( d = 2 \) and (ii) hybrid with critical exponent \( \zeta = 1/2 \) if \( d > 2 \) (which is consistent with a previous work [84]). For hypergraphs where both the vertex degrees and hyperedge cardinality distributions are Poissonian, the
Figure 8.3: \((2, \max(r, 2))-\text{core phase diagram}\). Phase diagram of \((2, \max(r, 2))-\text{core percolation}\) on hypergraphs with Poisson vertex degree distributions. Black circles and black line represent the phase boundary of \(d\)-uniform hypergraphs and hypergraphs with Poisson hyperedge cardinality distribution, respectively.

\(2, \max(r, 2)\)-core percolation is (i) continuous with critical exponent \(\zeta = 2\) if \(d < \bar{d} = 1\); (ii) continuous with critical exponent \(\zeta = 1\) if \(d = \bar{d}\); and (iii) hybrid with critical exponent \(\zeta = 1/2\) if \(d > \bar{d}\). The same set of critical exponents was found for the heterogeneous-\(K\)-core [85].
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

8.2.2 \( K = 2 \) and \( S = 2 \)

In this section we study the 2, 2-core. A similar definition of removable hyperedges to this one was used in [86], in the context of the core obtained from GLR procedure when studying the vertex cover problem in uniform hypergraphs. In this case, equation (8.7) reduces to

\[
\delta = \sum_{r=1}^{\infty} Q_h(r) \alpha^{r-1}. \tag{8.15}
\]

The relative size of 2, 2-core can be calculated by considering the probability that a randomly chosen vertex is connected to at least two non-removable hyperedges and the probability that a degree-one vertex is connected to a hyperedge with less than \((r - 2)\) other degree-one vertices. This results in

\[
s_{2c} = \sum_{k=2}^{\infty} P_u(k) \sum_{l=2}^{k} \binom{k}{l} (1 - \delta)^l \delta^{k-l}. \tag{8.16}
\]

Equations 8.10 and 8.15 have the same critical point as equations 8.8 and 8.7. Therefore, for \((2, 2)\)-core we recover the result found in the graph case that both the \((2, 2)\)-core and the giant component emerge at the same critical point [44]. In this case, the phase transition is always continuous (see solid lines in Figure 8.A.2 c and d, and for the studied hyperedge cardinality and vertex degree distributions, we have \(\beta = 2\). The condition of percolation transition is still given by equation 8.48.
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

Figure 8.4: \((K, S)\)-cores as a function of c. The relative size of \((2, S)\)-core \(s_{2c}\) as a function of the mean degree \(c\) for hypergraphs with Poisson degree distributions. For \(S = \max r - 1, 2\) (a) and (b), and \(S = 2\), (b) and (d). (a) and (c) \(d\)-uniform hypergraphs with \(d = 2, 3, 4\). (b) and (d) hypergraphs with Poisson hyperedge cardinality distribution and \(d = 1/2, 1, 2, 3\).
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

8.3 Discussion

Hypergraphs are natural generalizations of graphs. The percolation problems on hypergraphs have much richer phenomena than on graphs. We show that the emergence of the hyperdegree 2, $S$-core strongly depends on the threshold $S$ (i.e., it can emerge as either a continuous or a hybrid phase transition with different critical exponents). The heterogeneity of vertex degree and the hyperedge cardinality distributions, which in principle could offer more interesting phenomena, is not fully explored here,

Calculations and simulations

8.A Critical exponents and critical points

In this section we compute the critical exponents and critical points for the different phase transitions present in the previous section.

8.A.1 Emergence of the giant component in hypergraphs-critical point

Let us start with equations 8.1 and 8.2. We can rewrite equation 8.3 as

\[ F_g = 0, \quad (8.17) \]
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

where we define

\[ F_g \equiv \sum_{k=1}^{\infty} Q_n(k) \psi^{k-1} - \mu. \]  \hfill (8.18)

At the critical point when \( \mu^* = 1 \) and

\[ \frac{\partial F_g}{\partial \mu} \bigg|_{\mu = \mu^* = 1} = 0, \]  \hfill (8.19)

we obtain

\[ \frac{d_2 - d_1}{d_1} \frac{c^*_2 - c^*}{c^*} = 1, \]  \hfill (8.20)

where \( c^* \) and \( c^*_2 \) are, the hyperdegree first and second moment, respectively, at the critical point.

8.A.2 Emergence of the giant component in hypergraphs - critical exponent

Let us consider a point around the critical point, such as, \( \mu = 1 - \zeta \) with \( \zeta = 0^+ \) and \( c = c^* + \chi \) with \( \chi = 0^+ \). We can define the function from equation 8.18 as a function of \( \mu \) and \( c \) (i.e. \( F_g(\mu, c) \)). By expanding \( F_g(\mu, c) \) around the point \( (\mu, c) = (1, c^*) \) and combining
it with the result from equation 8.19 at the critical point, we can rewrite equation 8.18 as

\[
\frac{\partial F_g(\mu, c)}{\partial c} \bigg|_{(1, c^*)} \chi + \frac{\partial^2 F_g(\mu, c)}{\partial^2 \mu} \bigg|_{(1, c^*)} \zeta^2 + \frac{\partial^2 F_g(\mu, c)}{\partial^2 c} \bigg|_{(1, c^*)} \chi^2 \quad (8.21)
\]

\[
-2 \frac{\partial^2 F_g(\mu, c)}{\partial \mu \partial c} \bigg|_{(1, c^*)} \chi \zeta + (\cdots) = 0.
\]

For \( \mu = 1 \),

\[
F_g'(1, c) = 1 - c/c = 0,
\]  

(8.22)

it follows that

\[
\frac{\partial^n F_g(\mu, c)}{\partial^n c} \bigg|_{(1, c^*)} = 0
\]  

(8.23)

for any positive integer \( n \). Let us assume that there are no diverging moments for the rank

or hypergraph distribution. We can truncate our expansion of \( F_g \) at order 2 for \( \zeta \) and \( \chi \).

This implies that

\[
\zeta \sim \chi.
\]  

(8.24)

Let us expand \( s_g \) as a function of \( \zeta \),

\[
s_g = s_g(\mu) \bigg|_{\mu=\mu^*} - \frac{\partial s_g(\mu)}{\partial \mu} \bigg|_{\mu=\mu^*} \zeta + \frac{\partial^2 s_g(\mu)}{\partial^2 \mu} \bigg|_{\mu=\mu^*} \zeta^2 + (\cdots). \quad (8.25)
\]

For \( \mu^* = 1 \), \( s_g(\mu^*) = 0 \). We obtain

\[
s_g \sim \chi,
\]  

(8.26)
8.A.3 \((K, S)\)-core

Let us start with equation 8.6. It is useful to define

\[
F_{2c} \equiv \sum_{k=1}^{\infty} Q_n(k) \delta^{k-1} - \alpha = 0
\]  

(8.27)

at the critical point

\[
\left. \frac{\partial F_{2c}}{\partial \alpha} \right|_{(\alpha, c_1) = (\alpha^*, c^*)} = 0.
\]  

(8.28)

8.A.4 \((2, S_{\text{max}})\)-core - critical point

For \(S = \max(r, 2)\), \(\delta\) is defined in equation 8.9. Combining equations 8.27, 8.28 and 8.9, we obtain

\[
1 = \sum_{k=1}^{\infty} \sum_{r=2}^{\infty} Q_n^*(k) Q_h(r)(k-1)(r-1)\delta^*k-2(1-\alpha^*)r-2,
\]  

(8.29)

where \(Q_n^*(k)\) is the connectivity distribution with the critical mean hyperdegree \(e^*\). \(\delta^*\) and \(\alpha^*\) are the values of \(\delta\) and \(\alpha\) at the critical point, respectively. If at \(\delta^* = \alpha^* = 1\) we have that

\[
\left. \frac{\partial^2 F_{2c}}{\partial^2 \alpha} \right|_{\alpha = 1} > 0,
\]  

(8.30)

then the phase transition is continuous. From this condition, we obtain

\[
Q_h(2) \frac{c_2^* - c^*}{e^*} = 1.
\]  

(8.31)
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

From equations 8.45 and using a chain rule to describe the relation between $\delta$ and $\alpha$, we obtain a condition for the phase transition to be continuous:

$$2Q_n(3)(c_2^* - e^*) - Q_n(2)(c_3^* - 3c_2^* + 2e^*) < 0. \tag{8.32}$$

Let us assume a Poisson distribution of hyperdegrees, in this case $c_2^* - c^* = e^{*2}$ and $c_3^* - 3c_2^* + 2e^* = c^{*2}$. Combine these relations which equations 8.32 and 8.31 reduces to

$$2Q_n(3) < Q_n(2). \tag{8.33}$$

If the ranks of a hypergraph follows a Poisson distribution, there is a continuous phase transition when $d_1 < 1$, witch corresponds to

$$d_1 < \bar{d}, \tag{8.34}$$

where $\bar{d} \equiv 1$.

8.A.5  $(2, \max(r, 2))$-core - critical exponent

As before, we assume that the moments of the rank and hyperdegree distribution do not diverge. Let us consider points around the critical point, such as $\alpha = \alpha^* - \zeta$ with $\zeta = 0^+$ and $c = c^* + \chi$ with $\chi = 0^+$. We can define the function from equation 8.27 as a function of $\alpha$ and $\chi$ (i.e. $F_{2e}(\alpha, c)$). By expanding $F_{g}(\alpha, \chi)$ around the point $(\alpha, c) =$
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

($\alpha^*, c^*$) and combining it with result from equation 8.28 at the critical point, we can rewrite equation 8.18 as

$$\frac{\partial F_{2c}(\alpha, c)}{\partial c} \bigg|_{(\alpha^*, c^*)} + \frac{\partial^2 F_{2c}(\alpha, c)}{\partial^2 \alpha} \bigg|_{(\alpha^*, c^*)} \zeta + \frac{\partial^2 F_{2c}(\alpha, c)}{\partial^2 c} \bigg|_{(\alpha^*, c^*)} \chi^2 - 2 \frac{\partial^2 F_{2c}(\alpha, c)}{\partial \alpha \partial c} \bigg|_{(\alpha^*, c^*)} \chi \zeta + (\cdots) = 0.$$  

(8.35)

For the continuous phase transition, $\alpha^* = 1$, this implies

$$\frac{\partial^n F_{2c}(\mu, c_1)}{\partial^n c} \bigg|_{(1, c^*)} = 0$$  

(8.36)

for any positive integer $n$. Thus,

$$\zeta \sim \chi.$$  

(8.37)

Let us expand $s_{2c}$ as a function of $\zeta$,

$$s_{2c} = s_{2c}(\mu) \big|_{\alpha = \alpha^*} - \frac{\partial s_{2c}(\alpha)}{\partial \alpha} \bigg|_{\alpha = \alpha^*} \zeta + \frac{\partial^2 s_{2c}(\alpha)}{\partial^2 \alpha} \bigg|_{\alpha = \alpha^*} \zeta^2 + (\cdots),$$  

(8.38)

For $\alpha^* = 1$, $s_{2c}(\alpha^*) = 0$ and

$$\frac{\partial s_{2c}(\alpha)}{\partial \alpha} \bigg|_{\alpha = \alpha^* = 1} = 0.$$  

(8.39)

We obtain

$$s_{2c} \sim \zeta^2 \Rightarrow s_{2c} \sim \chi^2.$$  

(8.40)
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

For $s_{2c}^* = 0$ and $\eta = 2$. For the discontinuous phase transition, $\alpha^* \neq 1$, we have

$$\left. \frac{\partial F_{2c}(\alpha, c)}{\partial c} \right|_{(1, c^*)} \neq 0, \quad (8.41)$$

$s_{2c}(\alpha) \neq 0$ and

$$\left. \frac{\partial s_{2c}(\alpha)}{\partial \alpha} \right|_{\alpha = \alpha^*} \neq 0. \quad (8.42)$$

It implies

$$\zeta \sim \chi^{1/2}, \quad (8.43)$$

and

$$s_{2c} - s_{2c}^* \sim \chi^{1/2}, \quad (8.44)$$

that is equivalent to equation 8.11 for $\eta = 1/2$. A discontinuous phase transition with a critical exponent smaller than one considered a hybrid phase transition.

At the critical mean rank, $d_1 = \bar{d}$, $\alpha^* = 1$, but

$$\left. \frac{\partial^2 F_{2c}}{\partial^2 \alpha} \right|_{\alpha = 1} = 0. \quad (8.45)$$

In this case we have to take into account terms of order 3 in the expansion of equation 8.35.

We obtain,

$$\zeta \sim \chi^{1/2}, \quad (8.46)$$

87
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

and,

\[ s_{2c} \sim \chi, \quad (8.47) \]

8.A.6 \ (2, 2)-core - critical point

For \( S = 2 \), \( \delta \) is define in equation \( 8.15 \). In this case the phase transition is always continuous. We can compute the critical point by combining equation \( 8.27 \) \( 8.15 \) and 14 at \( \alpha = \delta = 1 \), obtaining

\[ \frac{d_2 - d_1 c^* - c^*}{d_1} = 1, \quad (8.48) \]

8.A.7 \ 2, 2-core - critical exponent

As before we, assume that the moments of the rank and hyperdegree distribution do not diverge. Let us consider points around the critical point, such as, \( \alpha = 1 - \zeta \) with \( \zeta = 0^+ \) and \( c = c^* + \chi \) with \( \chi = 0^+ \). We can define the function from equations \( 8.27 \) as a function of \( \alpha \) and \( \chi \) (i.e. \( F_{2c}(\alpha, c) \)). By expanding \( F_{2c}(\alpha, \chi) \) around the point \( (\alpha, c) = (1, c^*) \), and combining it with result from equation \( 8.28 \) at the critical point, we can rewrite equation \( 8.18 \) as

\[ \frac{\partial^2 F_{2c}(\alpha, c)}{\partial^2 \alpha} \bigg|_{(1,c^*)} \zeta^2 - 2 \frac{\partial^2 F_{2c}(\alpha, c)}{\partial \alpha \partial c} \bigg|_{(1,c^*)} \chi \zeta + ( \cdots ) = 0. \quad (8.49) \]
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

Note that for this case the equation 8.36 is still valid. We obtain

\[ \zeta \sim \chi. \]  \hspace{1cm} (8.50)

By combining this equations with the fact that \( s_{2c}(\alpha^*) = 0 \) and the result from equation 8.39, we obtain

\[ s_{2c} \sim \chi^2, \]  \hspace{1cm} (8.51)

8.A.8 Simulations

We compare our analytical calculations with random graphs generated in the following ways.

8.A.8.1 Poisson-Uniform hypergraphs

Poisson-Uniform hypergraphs we generate random hypergraphs where all hyperedges have the same rank \( d \) and the degrees of the nodes follows a Possion distribution with average degree \( c_1 \). We can generate this graphs by creating \( N \) nodes and \( L \) hyperedges of size \( d \).

We then fill the hyperedges with random selected nodes from our list of nodes. In the end

\[ c = d \times L / N \]  \hspace{1cm} (8.52)
CHAPTER 8. PERCOLATIONS ON HYPERGRAPHS

8.A.8.2 Poisson-Poisson hypergraphs

In the Poisson-Poisson hypergraphs we generate random hypergraphs where all hyperedges have the same rank and the degrees(ranks) of the nodes(hyperedges) follows a Possion distribution with average degree(rank) \( c = d \). We can generate this graphs by creating \( N \) nodes and \( L \) hyperedges. We then select a random nodes and add him to a random selected hyperedge. In the end

\[
c = d \cdot L / N
\]  

(8.53)

8.A.8.3 Results

Figures 8.A.1 to 8.A.3 shows that numerical simulations and the analitical calculations agree. (a) Plot of the of 2-core relative size, \( s_{2c} \), with \( S = 2 \), as a function of the mean hyperdegree , \( c \), for a uniform-Poisson hypergraph obtain from simulation. (b) Plot of the of 2-core relative size, \( s_{2c} \), with \( S = 2 \), as a function of the mean hyperdegree, \( c \), for a Poisson-Poisson hypergraph obtain from simulation. In the simulations we used networks with a size of a size \( 10^4 \).
Figure 8.A.1: (a) Plot of the size of the giant component, $s_g$, as a function of the mean hyperdegree, $c$, for a uniform-Poisson hypergraph obtained from simulation. The black solid lines represent the size of the respective giant component size using a mean field approach. (b) Plot of the size of the giant component $s_g$, as a function of the mean hyperdegree, $c$, for a Poisson-Poisson hypergraph obtained from simulation. The black solid lines represent the size of the respective giant component size using a mean field approach. In the simulations we used networks with a size of a size $10^4$. 
Figure 8.A.2: (a) Plot of the of 2-core relative size, $s_{2c}$, with $S = \max r - 1, 2$, as a function of the mean hyperdegree, $c$, for a uniform-Poisson hypergraph obtained from simulation. (b) Plot of the of 2-core relative size, $s_{2c}$, with $S = \max r - 1, 2$, as a function of the mean hyperdegree, $c$, for a Poisson-Poisson hypergraph obtained from simulation. In the simulations we used networks with a size of a size $10^4$. 
Figure 8.A.3: (a) Plot of the of 2-core relative size, $s_{2c}$, with $S = 2$, as a function of the mean hyperdegree, $c$, for a uniform-Poisson hypergraph obtained from simulation. (b) Plot of the of 2-core relative size, $s_{2c}$, with $S = 2$, as a function of the mean hyperdegree, $c$, for a Poisson-Poisson hypergraph obtained from simulation. In the simulations we used networks with a size of a size $10^4$. 

93
Greedy Leaf Removal on Hypergraphs

The core of a graph – defined as the remainder of the greedy leaf removal (GLR) procedure where leaves (vertices of degree one) and their neighbors are removed iteratively from the graph – has been related to the conductor-insulator transition [12] structural controllability [13] and many combinatorial optimization problems [14]. In fact, the size of the core is related to a fundamental combinatorial problem: the complexity of the minimum edge cover problem, being either solvable problem in polynomial time if the relative size of the core is zero, or generally a NP-hard problem if the relative size of the core is non-zero [15][16]. In contrast, the minimum edge cover problem on graphs can be computed in polynomial time [87], yet, this is not true for general hypergraphs [88]. These two problems are relevant to many real-world problems in different fields, e.g. finding the optimal drug combination in medicine and pharmacology [89], or searching for files in a storage system [90]. For percolation of hypergraphs in regular lattices such as square lattices, see [91][92][93][94].

We can generalize the GLR procedure to the hypergraph case in two slightly different
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

ways, one associated with the edge-cover problem and another associated with the vertex-cover problem.

Let’s consider the edge-cover problem and the hypergraph in Figure 9.1. $e_2$ covers the same set of nodes as $e_1$, but also covers node $v_1$, thus $e_1$ is not necessary (Figure 9.1(c)). Similarly, because $v_2$ is contained by all hyperedges that contain node $v_1$, if $v_1$ is covered, node $v_2$ is also covered; thus, we only need to cover $v_1$ and we know that $v_2$ is also being covered (Figure 9.1(e)). The vertex-cover problem is the dual problem to the edge-cover problem, so we obtain the exact opposite set of rules. In Figure 9.1(f) we can see that, in the context of the vertex-cover problem, node $v_1$ is redundant because $v_2$ covers the same hyperedge as $v_1$ and also covers hyperedge $e_1$. In the same way, if node $v_1$ is covered, node $v_2$ is also covered because $v_2$ contains all nodes that contain $v_1$ (Figure 9.1(d)).

Let’s define three sets of hyperedges (nodes): a set of hyperedges (nodes), $S$, that is a solution for the minimum edge (vertex)-cover problem; a set of hyperedges (nodes) $\tilde{S}$, containing hyperedges (nodes) that can determined to be part $S$ using our approach, and $core_1$core_2 containing the nodes we cannot be determined if they belong to $S$ or not using our approach. If the degree (cardinality) of a node (hyperedge) is zero, then that node (hyperedge) can not be covered. For the remaining nodes (hyperedges) we can find the nodes that belong to each category using the following procedure:

- **Rule 1**: Let’s consider hyperedges $e_1$ and $e_2$, which contain a set of nodes $N_1$ and $N_2$. If $N_1 \subseteq N_2$ we remove $N_1(N_2)$ from the network.

- **Rule 2**: Let’s consider nodes $n_1$ and $n_2$ that are contained by a non empty set of
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

Figure 9.1: GLR core. (a) Example of the generalized greedy leaf removal for the minimum edge, ((a),(c),(e),(g) and (i)), and vertex cover problem ((b),(d),(f),(h) and (l)). Dots represent nodes and circles represent hyperedges. The green nodes (hyperedges) in (a) and (b) are a solution to the minimum edge cover set. Dashed circles represent redundant hyperedges.

If $E_1 \subseteq E_2$ we remove $E_2(E1)$ from the network.

We repeat this process until no more nodes can be removed. In the final network,

- Every hyperedge(node) with cardinalty (degree) one belongs to $\tilde{S}$ (Figure 9.1(i) and (l)).
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

- Every hyperedge(node) with cardinalty(degree) bigger than one belongs to core1(core2).

It is important to notice that these two algorithms reduces to the standard greedy leaf removal algorithm to solve the vertex-cover problem in graphs if all hyperedges have cardinality two.

9.1 The core of random hypergraphs

In this section we study the core of two random hypergraphs, Poisson-Uniform and Poisson-Poisson hypergraphs,

- In the Poisson-Uniform hypergraphs we generate random hypergraphs where all hyperedges have the same rank \(d\) and the degrees of the nodes follows a Possion distribution with average degree \(c_1\).

- In the Poisson-Poisson hypergraphs we generate random hypergraphs where all hyperedges have the same rank and the degrees(ranks) of the nodes(hyperedges) follows a Possion distribution with average degree(rank) \(c(d)\).

In both of these hypergraphs the probability of two hyperedges or nodes being connected to the same two or more distinct hyperedges or nodes, respectively, is very small. Therefore, the only way rules 1 and 2 apply if nodes or hyperedges with degree or cardinality one are involved.
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

9.1.1 core1

For these networks, core1 can be computed using the following procedure. First, we re-
move all hyperedges with cardinality one, hyperedges that contain vertices with degree one
(called leaves) vertices contained in those hyperedges. Note that some of the vertices in
those hyperedges may have degrees larger than one. After we remove those vertices, the
cardinality of other hyperedges that contain those vertices will decrease, and if the hyper-
edge cardinality drops below two, we remove those hyperedges. If eventually there is no
core left, the difference between approximate solution from this GLR procedure and the
exact solution is zero in the thermodynamic limit (see section 9.A for details).

To study the core percolation on hypergraphs, we generalize the mean-field approach
proposed for the graph case [26]. We define two types of removable vertices: a vertex is
(i) $\alpha$-removable if it is or can become a vertex with a degree one and (ii) $\beta$-removable if its
degree is larger than one and belongs to at least one leaf hyperedge. Dually, we define two
types of removable hyperedges: a hyperedge is (i) $\delta$-removable if it is or can become an leaf
hyperedge; (ii) $\epsilon$-removable if it has cardinality $r$ and is removed because it is connected
to $(r - 1) \beta$-removable vertices. Consider a large uncorrelated random hypergraph $\mathcal{H}$
with arbitrary vertex degree and hyperedge cardinality distributions. We can determine the
category of a vertex $v$ in $\mathcal{H}$ by the categories of its neighboring hyperedges in the modified
hypergraph $\mathcal{H} \setminus v$ with vertex $v$ and all its hyperedges removed from $\mathcal{H}$. To determine
this we use the following rules: (i) $\alpha$-removable vertex: all neighboring hyperedges are
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

$\epsilon$-removable; (ii) $\beta$-removable vertex: at least one neighboring hyperedge is $\delta$-removable. Similarly, we can determine the category of a hyperedge $e$ in $\mathcal{H}$ by the categories of its neighboring vertices in the modified hypergraph $\mathcal{H} \setminus e$ with hyperedge $e$ and all its vertices removed from $\mathcal{H}$, using the following rules: (iii) $\delta$-removable hyperedge: at least one neighboring vertex is $\alpha$-removable; (iv) $\epsilon$-removable hyperedge: at least one neighboring vertex is $\beta$-removable. Let $\alpha$ (or $\beta$) denote the probability that a random neighboring vertex of a random hyperedge $e$ in a hypergraph $\mathcal{H}$ is $\alpha$-removable (or $\beta$-removable) in $\mathcal{H} \setminus e$. Let $\delta$ (or $\epsilon$) denote the probability that a random neighbor of a random vertex $v$ in a hypergraph $\mathcal{H}$ is $\alpha$-removable (or $\beta$-removable) in $\mathcal{H} \setminus v$. Then rules (i)-(iv) enable us to derive a set of self-consistent equations:

\[
\alpha = \sum_{k=1}^{\infty} Q_n(k) \epsilon^{k-1},
\]

\[
1 - \beta = \sum_{k=1}^{\infty} Q_n(k)(1 - \delta)^{k-1},
\]

\[
1 - \delta = \sum_{r=1}^{\infty} Q_h(r)(1 - \alpha)^{r-1},
\]

\[
\epsilon = \sum_{r=1}^{\infty} Q_h(r)\beta^{r-1}.
\]

The relative size of core 1 is given by

\[
s_{core1} = \sum_{k=2}^{\infty} P_n(k) \sum_{l=2}^{k} \binom{k}{l} (1 - \delta - \epsilon)^l \epsilon^{k-l}.
\]

For hypergraphs with a Poisson vertex degree distribution and different hyperedge car-
dinality distributions, we find that the core emerges as a continuous phase transition (see Figure 9.1),

\[ s_{\text{core1}} \propto (c - c^*)^{\zeta} \]  

(9.6)

with a critical exponent \( \zeta = 1 \) (see section 9.A for details). The relation between the critical mean degree \( c^* \) (percolation threshold) and the hyperedge mean cardinality \( d \) is represented in figure 9.1.

9.1.2 core2

The core2 applied to Poisson-Uniform and Poisson-Poisson hypergraphs can be computed using the following procedure. First, we remove all the vertices with degree one and all the vertices connected to hyperedges with cardinality one together with all the hyperedges connected to those vertices. Note that some of the hyperedges connected to those vertices may have cardinality larger than one. After we remove those hyperedges (defined as leaves), the degree of the other vertices connected those hyperedges will decrease, and if the vertex degree drops below two, we remove those vertices. If the core size is non-zero, the vertex cover problem is, in general, an NP-hard problem [16].

There are two types of removable vertices: (i) \( \alpha \)-removable vertices are vertices of degree one; (ii) \( \beta \)-removable vertices are vertices with degree larger than one that belong to at least one leaf hyperedge. Similarly, there are two types of removable hyperedges: (i) \( \delta \)-removable hyperedges are hyperedges that are or can become leaf hyperedges; and (ii) \( \epsilon \)-removable hyperedges are hyperedges that are removed because they are connected

100
Figure 9.1: **GLR cores as a function of c.** The relative core size of $s_{\text{core1}}$ and $s_{\text{core1}}$ for hypergraphs with Poisson degree distributions. (a) $s_{c1}$ for $d$-uniform hypergraphs with $d = 2, 3, 4$. $s_{c1}$, (b) and $s_{c2}$ and (d) for hypergraphs with Poisson hyperedge cardinality distribution, and $d$-uniform hypergraphs for different values of $d$.

to at least one $\beta$-removable vertex. The vertices obey the same relations as before (i.e., Equations 9.1 and (refcore:eq2), and we derive the following self-consistent equations for
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

the hyperedges:

\[ \delta = \sum_{r=1}^{\infty} Q_h(r) \alpha^{r-1}, \quad (9.7) \]

\[ 1 - \epsilon = \sum_{r=1}^{\infty} Q_h(r) (1 - \beta)^{r-1}. \quad (9.8) \]

There is another way to obtain the same result. The core2 of a hypergraph is the core1 of the dual hypergraph (i.e. a hypergraph whose vertices and hyperedges are interchanged). Thus, we can obtain the same results shown in equation (9.1) (9.4) by the following transformation:

\[ \alpha \rightarrow \delta, \]

\[ \delta \rightarrow \beta, \]

\[ \beta \rightarrow \epsilon, \]

\[ \epsilon \rightarrow \alpha, \]

\[ Q_h(r) \rightleftharpoons Q_n(r). \]

The relative size of the core \( s_c \) is given by the number of vertices connected to at least two non-removable hyperedges. Hence, we obtain

\[ s_{\text{core2}} = \sum_{k=2}^{\infty} P_n(k) \sum_{l=2}^{k} \binom{k}{l} (1 - \delta - \epsilon)^l \epsilon^{k-l}. \quad (9.9) \]
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

For hypergraphs with a Poisson vertex degree distribution and different hyperedge cardinality distributions, we find that the core emerges as a continuous phase transition (see Figure [9.1])

\[ s_{\text{core2}} \sim (c - c^*)^\zeta, \]  

(9.10)

with a critical exponent \( \zeta = 1 \). Figure [9.1](d) shows that for a Poisson-Poisson hypergraph the size of core2 starts to decrease at large values of \( c \). By increasing the number of hyperlinks connected to a node, but keeping the cardinality distribution constant, the probability of a node being connected to a hyperedge with cardinality one increases, and any node connected to a hyperedge with cardinality one is automatically removed. This effect is not relevant if the probability that a node is connected to a hyperedge with cardinality one is very small, \( 1 - \exp(-ce^{-d}) \ll 1 \). For large values of \( c \) and \( d \), this effect is only relevant if \( c \sim \exp(d) \).

For \( d \)-uniform hypergraphs with Poisson vertex degree distribution, there is a simple relation between the critical mean degree (percolation threshold) and the mean hyperedge cardinality:

\[ c^* = \frac{e}{d - 1}, \]  

(9.11)

where \( e = 2.71828 \cdots \). This result was previously found in [86] using an algorithm design for uniform hypergraphs. Figure [9.1] shows relation between the critical mean degree \( c^* \) (percolation threshold) and the hyperedge mean cardinality \( d \). The phase space of core2 is equal to the phase space of core1 if we interchanged the mean cardinality \( d \) with the mean
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

![Figure 9.2: GLR cores phase space. Phase diagram of the core percolation on hypergraphs with Poisson vertex degree distributions. Black circles and black lines represent the phase boundary of $d$-uniform hypergraphs and hypergraphs with Poisson hyperedge cardinality distribution, respectively. (a) core1. (b) core2.]

degree $c$. Of course this is true because as mentioned before, the core2 of a hypergraph is the core1 of the dual hypergraph.

### 9.2 Application

Because real-world hypergraphs tend to have specific properties not captured by the random models we have studied before, in this section we compute the core of real-word hypergraphs using rules 1 and 2. We find that, because the cores are very small, the edge and vertex cover problems are effectively solvable in those hypergraphs. We apply this procedure to three data sets.
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

- APS data set is composed of the articles published in all APS journals from 1893 to 2010. We take the authors and articles as the nodes and edges of the network, respectively.

- Drug-target interaction network from the DrugBank [95]; it is composed by the set of drugs and the genes with each each drug interacts with. We take the genes and drugs as the nodes and edges of the network, respectively.

- Genome-scale metabolic network of E. coli; it is composed by the metabolic network of E. coli obtained from the BiGG project [96]. Each reaction is treated as a hyperedge and the metabolites involved in a reaction are the nodes.

Figure 9.1(a) shows that the size of core1(core2) is 3.4% (81 edges) for core1 for the genome-scale metabolic network and less than 0.2% for all other cores. Figure 9.1(b) shows the fraction of nodes(edges) necessary to cover all the nodes(edges), upper and lower bounds. Because core1(core2) is very small, the upper and lower are very close to each other, making the edge and vertex cover problem effectively solvable in these case. We also compare the size of each core with two random versions of the network. In random1, we use the degree cardinality distribution from each network to compute the size of the core using equations 9.9 and 9.5 respectively. In random2, we consider a random Poisson-Poisson hypergraph with the same average degree and cardinality as the original network. For all random1 hypergraphs, the size of the core is always zero and in random2, hypergraph the size of the core is between 30% and 100 % of the entire network. The size of the actual
core of a network is slightly bigger than or equal to random1, and it is much smaller than random2, indicating that the degree and cardinality distributions are potentially the main factors that explain the small cores of these hypergraphs.
9.3 Discussion

The edge and vertex cover problems are related to many other optimization problems (e.g., finding the optimal drug combination in medicine and pharmacology [89], or searching for files in a storage system [90]). Our approach might offer a new set of tools for solving classical NP-hard problems in a real world system in polynomial time. It also opens up some important questions. Why are these cores small in real-world hypergraphs, especially in biological networks? Are there other classes of problems in biological systems which are usually NP-hard but can be solved? In particular, metabolic and protein-protein networks are problems that play an important role in regulating/controlling or optimizing the cell. It stands to reason that most optimization inside a cell or other biological system is done locally through local or greedy rules. Therefore, it would make sense that cells are optimized to make these problems easy to solve. Understanding if this class of problems exists and how they can be solved would lead to a deeper comprehension of how biological systems work.

Calculations and simulations

In this section we compute the critical exponents and critical points for the different phase transitions present in the previous section.
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

9.A    Critical exponents and critical points

9.A.1   Core1 for uniform hypergraphs - critical point

There is a simple relation for the core1 critical point and the mean rank $d_1$ for uniform hypergraphs with $d_1 > 1$. Let us start by rewriting Equation 9.5 as

$$ F_{c1}(\alpha, c) = 0 $$  (9.12)

where we define,

$$ F_{c1}(\alpha, c) \equiv \sum_{k=1}^{\infty} Q_n(k) e^{k^{-1} - \alpha}. $$  (9.13)

For uniform-Poisson hypergraph equations 9.1 to 9.4 can be reduced to

$$ x = e^{-c(d_1-1)y} $$  (9.14)

$$ y = e^{-c(d_1-1)x}, $$  (9.15)

where we define $x \equiv \alpha^{r-1}$ and $y \equiv (1 - \beta)^{r-1}$. The function defined in equation 9.13 can be written as a function of $x$ and $c$,

$$ F_{c1}(x, c) \equiv \exp \left( -c \left( d_1 - 1 \right) e^{-c \left( d_1 - 1 \right) x} \right) - x. $$  (9.16)

The critical point is given by $\alpha^* = \left( e^* \left( d_1 - 1 \right) \right)^{1/(1-d_1)}$ and $c^* = e/(d_1 - 1)$. 

CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

9.A.2 Core1 - critical exponent

Equations 9.1 to 9.4 can be written as

\[ \alpha = A(1 - z, c) \]  \hspace{1cm} (9.17)
\[ z = B(h) \]  \hspace{1cm} (9.18)
\[ h = A(1 - \delta, c) \]  \hspace{1cm} (9.19)
\[ \delta = B(\alpha) \]  \hspace{1cm} (9.20)

where we define \( z \equiv 1 - \epsilon, h \equiv 1 - \beta, A(1 - z) \equiv \sum_{k=1}^{\infty} Q_n(k)(1 - z)^{k-1} \) and 
\( B(h) \equiv \sum_{r=2}^{+\infty} Q_r(r)h^{r-1} \). \( F_{c1}(\alpha, c) \) can be written as

\[ F_{c1}(\alpha, c) = H(H(\alpha, c), c) - \alpha, \]  \hspace{1cm} (9.21)

where we define

\[ H(\alpha, c) \equiv A(1 - B(\alpha), c). \]  \hspace{1cm} (9.22)

This function has the same form as a similar equation for the graph case, studied in eq S60 from [26]. The only difference is the way we define the function \( H(\alpha, c) \). Nevertheless it obeys the same relation, and therefore has the same critical exponent \( \zeta_1 = 1 \), i.e:

\[ \left. \frac{\partial H(\alpha, c)}{\partial \alpha} \right|_{(\alpha, c) = (\alpha^*, c^*)} = -1, \]  \hspace{1cm} (9.23)
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

\[
\left. \frac{\partial F_{c2}(\alpha, c)}{\partial c} \right|_{(\alpha, c) = (\alpha^*, c^*)} = 0, \tag{9.24}
\]

\[
\left. \frac{\partial^2 F_{c2}(\alpha, c)}{\partial^2 \alpha} \right|_{(\alpha, c) = (\alpha^*, c^*)} = 0. \tag{9.25}
\]

From an expansion of \( F_{c1} \) as a function of \( (c - c^*) \), for \( c - c^* = 0^+ \), it follows that

\[
\alpha - \alpha^* \sim (c - c^*)^{1/2}, \tag{9.26}
\]

\[
\beta - \beta^* \sim (c - c^*)^{1/2}, \tag{9.27}
\]

\[
1 - \beta - \alpha \sim (c - c^*)^{1/2}, \tag{9.28}
\]

and,

\[
s_{c1} \sim (c - c^*), \tag{9.29}
\]

which is is equivalent to equation 9.6, for \( \zeta_1 = 1 \). We confirm this result using the numerical solution of equation 9.1 to 9.4.
9.A.3 Core2 - critical exponent

Equations 9.1, 9.2, 9.7 and 9.8 can be written as

\[ \alpha = A(\epsilon, c) \] (9.30)

\[ \epsilon = B(1 - u) \] (9.31)

\[ u = A(v, c) \] (9.32)

\[ v = B(1 - \alpha) \] (9.33)

where we define \( u \equiv 1 - \beta \) and \( v \equiv 1 - \delta \). The function \( F_{c2}(\alpha, c) \equiv A(\epsilon, c) - \alpha \) can be written as

\[ F_{c2}(\alpha, c) = G(G(\alpha, c), c) - \alpha, \] (9.34)

where we define

\[ G(\alpha, c) \equiv A(B(1 - u), c1). \] (9.35)

This function has the same form as equation 9.22, the only difference is how we define the function \( G(\alpha, c) \). Nevertheless it obeys the same relations as equations 9.23 to 9.28, thus,

\[ s_{c2} \sim (c - c^*)^{\zeta_2}, \] (9.36)

with \( \zeta_2 = \zeta_1 = 1 \). We confirm this result using the numerical solution of equations 9.1, 9.2, 9.7 and 9.8.
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

9.1.4 Simulations

We compare our analytical calculations with random graphs generated in the following ways.

9.1.4.1 Poisson-Uniform hypergraphs

For Poisson-Uniform hypergraphs, we generate random hypergraphs where all hyperedges have the same rank \( d \) and the degrees of nodes follow a Possion distribution with average degree of \( c_1 \). We can generate this graphs by creating \( N \) nodes and \( L \) hyperedges of size \( d \). We then fill the hyperedges with random selected nodes from our list of nodes. In the end

\[
c = d \times L/N
\]  \hspace{1cm} (9.37)

9.1.4.2 Poisson-Poisson hypergraphs

For Poisson-Poisson hypergraphs we generate random hypergraphs where all hyperedges have the same rank and the degrees(ranks) of the nodes(hyperedges) follows a Possion distribution with an average degree(rank) \( c(d) \). We can generate these hypergraphs by creating \( N \) nodes and \( L \) hyperedges. Then we select random nodes and add them to a randomly selected hyperedge. In the end,

\[
c = d \times L/N
\]  \hspace{1cm} (9.38)
9.1.4.3 Results

Figures 9.1 to 9.2 show that the numerical simulations and the analytical calculations agree.

Figure 9.1: Core1 obtained from simulations. (a) Plot of the core1 relative size, $s_{c1}$, as a function of the mean hyperdegree, $c$, for a uniform-Poisson hypergraph, obtained from simulations. Red circles represent the mean rank, $d_1 = 2$; blue circles $d_1 = 3$; green circles $d_1 = 4$, and black solid lines represent the size of the core obtained from the mean field approach. (b) Plot of the core1 relative size, $s_{c1}$, as a function of the mean hyperdegree, $c$, for a Poisson-Poisson hypergraph. It was obtained from simulations. Red circles represent the mean rank, $d_1 = 1$; blue circles $d_1 = 2$; green circles $d_1 = 3$, and black solid lines represent the size of the core obtained from the mean field approach. In the simulations we used networks which a size of $10^5$. 

113
CHAPTER 9. GREEDY LEAF REMOVAL ON HYPERGRAPHS

Figure 9.2: Core2 obtained from simulations. (a) Plot of the size the giant component, $s_{c1}$, as a function of the mean hyperdegree, $c$, for a uniform-Poisson hypergraph obtained from simulations. Red circles represent the mean rank, $d_1 = 2$; blue circles $d_1 = 3$; green circles $d_1 = 4$; black solid lines represent the size of the core obtained using a mean field approach. (b) Plot of the core2 relative size, $s_{c2}$, as a function of the mean hyperdegree, $c$, for a Poisson-Poisson hypergraph obtained from simulations. Red circles represent the mean rank, $d_1 = 2$; blue circles $d_1 = 3$; green circles $d_1 = 4$, and black solid lines represent the size of the core obtained using a mean field approach. In the simulations we used networks which a size of $10^5$. 
Conclusion

We showed how network science can be used to study the cosmic web. Our results represent only the first step towards a network-based understanding of the universe, but it offers a framework for the systematic study the cosmic web. We also explored a hypothetical spreading process at the cosmic level. We found that the structure of the universe makes this process non-trivial, and we found a very peculiar phase transition, resembling Griffith phase transitions [75]. The combination of the expansion of the universe with the structure of universe makes the spreading process localized to a small region of the universe. In the second part of our work, we explored a simple generalization of graphs, hypergraphs that provide a more faithful representation of many real networks. We found that in all tested real hypergraphs, two key combinatorial problems, the vertex and edge cover problems could be solved effectively solved in polynomial time. While the traditional metric to measure the computational complexity theory of a problem is extremely useful, it always deals in the worst case possible, and the worst case possible is not necessarily typical or even relevant to real world problems. Our results suggest that the edge and vertex cover
CHAPTER 10. CONCLUSION

problems might not be hard problems at all in real cases associated with sociology and biology. This result might not only impact the way people compute approximated solutions for edge and vertex cover problems, and related problems, but also opens a new research path for algorithms based on the minimum vertex and edge cover set for possible clustering, hierarchical clustering, and community detection based on minimum vertex and edge cover.

In conclusion, we showed how graphs can be used to gain insight from complex systems, and we showed that a small generalization of graph called hypergraphs opens new paths for a better and richer representation of real complex systems. Counter-intuitively, that patch may make solving complex problems easier in real-world hypergraphs than in graphs.
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


[64] Juhan Kim, Changbom Park, and Yun-Young Choi. A subhalo-galaxy correspondence
BIBLIOGRAPHY


Mass Assembly (GAMA): the red fraction and radial distribution of satellite galaxies. 


BIBLIOGRAPHY


[83] Béla Bollobás and Oliver Riordan. Erdős Centennial, chapter The Phase Transition in


