Random Preferential Attachment Hypergraph With Vertex Deactivation

by

Kostiantyn Ohulchanskyi

Supervised by

Frédéric Giroire
Nicolas Nisse
Małgorzata Sulkowska
Thibaud Trolliet

Université Côte d'Azur
Polytech Nice Sophia

August 2021

Abstract

Collaboration networks can be naturally represented as hypergraphs, where vertices correspond to authors and hyperedges correspond to publications. In the past decades, several random hypergraph models have been proposed. However, these models do not generate power-law degree distributions with an exponential cutoff, which we observe in real-world collaboration networks. We present a novel mathematical model, which utilizes the preferential attachment and the vertex deactivation mechanisms to generate random hypergraphs. We prove that, under a few assumptions, the degree distribution of such random hypergraphs follows a power-law distribution with an exponential cutoff. Finally, we show simulation results, which demonstrate the close correspondence between the degree distribution of a hypergraph generated according to the proposed model and a theoretical power-law distribution with an exponential cutoff.

Contents

1 Introduction

The field of complex networks investigates structures with complex topology. Many real-world systems can be described as complex networks: consider, for example, the World Wide Web, which consists of web pages and hyperlinks, or social networks, which consist of people and connections between them.

In the past decades, numerous researchers analysed empirical data to study different features of complex networks [5, 11, 14, 19]. In particular, a lot of research has been conducted to estimate degree distributions of various real networks and to develop mathematical models, which would explain how these degree distributions could be generated. It is often believed that many real-world networks have power-law degree distributions, which is why numerous existing models aim to generate random networks whose degree distribution follows a power law [2, 3, 10]. However, several studies of collaboration networks show that their degree distributions rather follow a power-law distribution with an exponential cutoff [23–25]. This renders the existing models unsuitable for modelling collaboration networks.

1.1 Motivation

Initiatives d'Excellence (IDEX) and Laboratoires d'Excellence (LABEX) are French funding programs that aim to promote scientific collaborations between different countries, universities, researchers, research teams and disciplines by providing funds for interdisciplinary research. The main goal of these programs is to increase the productivity of researchers and reinforce the partnership between institutions and teams.

This thesis is a part of the Scientific Networks and IDEX Funding (SNIF) project conducted by Inria Sophia Antipolis, I3S, GREDEG and SKEMA Business School. The purpose of the SNIF project is to measure the success of the aforementioned funding programs by studying the impact of funding on the evolution of the collaboration network.

To the best of our knowledge, there are no existing mathematical models, which generate random hypergraphs whose degree distribution follows a power-law distribution with an exponential cutoff. As such models would be useful for modelling collaboration networks, in this thesis, we aim to develop a novel mathematical model producing such random hypergraphs.

1.2 Contributions

We present a novel mathematical model, which combines the existing and well-studied growth and preferential attachment mechanisms with the vertex deactivation mechanism to generate random hypergraphs that are suitable for modelling collaboration networks. Then, we conduct a deep analysis of the proposed model and, under a few assumptions, prove a theorem, stating that the degree distribution of a generated hypergraph follows a power-law distribution with an exponential cutoff. At last, we also discuss a method to precisely calculate the parameters of this theoretical distribution.

1.3 Thesis Organization

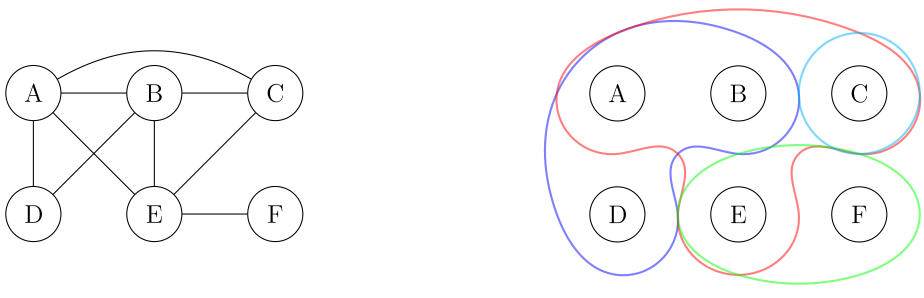
The thesis is organized as follows.

- In Chapter 2, we discuss the background of complex networks and mention the most influential related works in the field.
- In Chapter 3, we formally define the random hypergraph model. Also, we state and prove several theorems, related to the degree distribution of hypergraphs generated according to the model.
- In Chapter 4, we discuss hypergraphs that were simulated according to the model.
- Finally, we conclude the thesis with Chapter 5.

2 Background

Complex networks can naturally be represented as graphs or, in a more general case, *hypergraphs*. Such a representation enables us to apply mathematical tools in order to study the properties of these networks, such as the degree distribution, the structure of communities, distances between participants, and many other metrics. Another method that can be applied to study complex networks is random models, which often define a random process, describing how a graph or a hypergraph evolves over time. In this way, we can approximate the evolution of a real network by specifying a set of simple rules, which can later be rigorously analysed.

In this thesis, we focus on the analysis of *collaboration networks*. There are several ways to represent a collaboration network. The first being a *collaboration graph*, where each vertex represents an author, and there is an edge between two authors if they collaborated on a publication. However, one of the limitations of this representation is that we lose all the information about publications (neither do we know how many publications each author published, nor how publications are distributed between authors; we also lose information about publications that were published by a single author). In order to overcome this problem, one could consider other representations of the network, such as weighted or bipartite graphs, or *collaboration hypergraphs*. In a collaboration hypergraph, each hyperedge represents a publication, connecting co-authors together.



(a) (b)

Figure 2.1: Different representations of a collaboration network, consisting of 6 authors and the following publications: $\{A, B, C, E\}$, $\{A, B, D\}$, $\{E, F\}$ and $\{C\}$. A publication corresponds to a clique in a collaboration graph and a hyperedge in a collaboration hypergraph.

2.1 Power Laws

The degree distribution of a network, which is defined as

$$P(k) = \frac{\text{number of vertices of degree } k}{\text{number of all vertices}}, \tag{2.1}$$

can give us a good preliminary insight into the structure of the network. It is often believed that most real-world networks are *scale-free*, meaning that their degree distribution follows a *power-law distribution* [21, 26] of the form

$$P(k) \sim Ck^{-\alpha}, \tag{2.2}$$

where $\alpha > 0$ is the exponent parameter and $C > 0$ is the scaling constant.[1] In the past years, a lot of research has been conducted to establish scale-free properties of complex networks. The power-law distributions have been observed, for example, in citation networks [11], software dependency networks [19], the Internet [14] and others.

However, more recent studies [6, 28] question the ubiquitousness of power laws and show that scale-free networks do not appear as often as previously expected. More importantly, an investigation of a real-world collaboration network confirmed that a power-law distribution is a poor fit to the degree distribution, meaning that the network that was being investigated was not scale-free [22–25]. Instead, it was shown that a *power-law distribution with an exponential cutoff* of the form

$$P(k) \sim Ck^{-\alpha}\gamma^k, \tag{2.3}$$

where $0 < \gamma \leq 1$ is a constant parameter of the distribution, is a significantly better fit. Furthermore, we confirmed this observation by analysing a different real-world collaboration network, based on data collected from the Scopus database [27]; we discuss this analysis further in Section 2.2.

Note that by letting $\gamma = 1$, we obtain the "pure" power-law distribution. Hence, the power-law distributions with an exponential cutoff define a broader family of distributions, which implies that it will always be a better fit. Therefore, it is important to understand whether the difference between these distributions is significant to see if a power-law distribution is a good fit.

[1] For two real functions $f(k)$ and $g(k)$, we denote $f \sim g$ if $f(k)/g(k) \to 1$ as $k \to \infty$.

2.2 Analysis of a Real-World Network

We studied a real collaboration network, based on 239 414 publications in computer science between the years 1990 and 2018 that were extracted from the Scopus database [27]. The corresponding collaboration graph contained 258 145 vertices and 1 849 527 edges. Our main goal was to establish whether the constructed collaboration graph and hypergraph were scale-free or not. In order to check that, we applied state-of-the-art statistical tools [9] to fit and compare different theoretical distributions, including the power-law distribution and the power-law distribution with an exponential cutoff.

As a result, it turned out that neither the collaboration graph, nor the collaboration hypergraph was scale-free. In fact, the difference between the fits of the power-law distribution and the power-law distribution with an exponential cutoff was statistically significant. Figure 2.2 demonstrates the difference between the empirical degree distribution and the two fits.

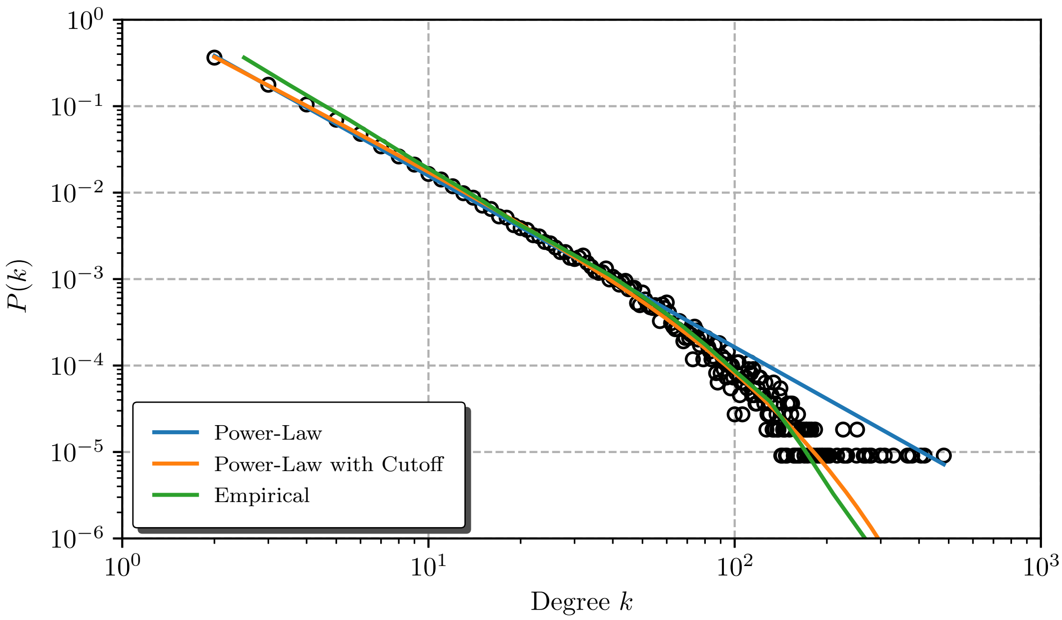


Figure 2.2: Comparison of the fit of different theoretical distributions to the degree distribution of a real-world collaboration network. From the figure, it can be seen that the difference between the fits of the power-law distribution and the power-law distribution with an exponential cutoff is significant. For example, one of the necessary (but not sufficient) conditions for a distribution to follow a power-law distribution is that it must appear as a straight line on a log-log scale plot. However, as can be seen above, the empirical degree distribution does not seem to be a straight line, and we observe the presence of a cutoff in its tail.

2.3 Related Works

In the last 50 years, numerous random graph models have been proposed. To name a few: the Erdős–Rényi model [12, 13], which uniformly selects a graph from a set of all possible graphs of a certain size; the Barabási–Albert model [3], which generates scale-free graphs using the growth and preferential attachment mechanisms; the Cooper et al. model [10], which generates scale-free graphs similarly, but also considers the vertex deletion; the Chung–Lu model [7], which generates graphs with the expected degree sequences; and many others.

Also, a few models that generate random hypergraphs have been developed. The first work in this field was conducted by Bollobás and Erdős [4], and it aimed to investigate cliques in random graphs. Later, another work by Ghoshal et al. [17] studied how random hypergraph models can be applied to analyse complex networks. In 2010, Wang et al. [29] published the first random hypergraph model, which used the preferential attachment mechanism. Recently, Avin et al. [2] proposed a mathematical model, which generates random scale-free hypergraphs using the growth and preferential attachment mechanisms, similarly to the Barabási–Albert model.

However, as we discussed in the previous sections, it appears that the real-world collaboration network are not scale-free, and their degree distribution rather follows a power-law distribution with an exponential cutoff. Currently, only several models that generate such distributions have been proposed [15, 16]. These models also utilize the growth and preferential attachment mechanisms, but they introduce the vertex deletion and deactivation mechanisms to make the exponential cutoff appear. Nevertheless, the problem is that these models are not described in terms of graphs and hypergraphs, and they cannot be directly applied to generate random collaboration networks. Hence, we aim to adapt the ideas discussed in [16] to develop a mathematical model, which generates random hypergraphs whose degree distribution follows a power-law distribution with an exponential cutoff.

3 Model

Many mathematical models, which generate scale-free graphs or hypergraphs, utilize the growth and preferential attachment mechanisms [2, 3] to achieve a power-law degree distribution. Informally, by growth, we mean that some evolution of the initial hypergraph is involved: that is, we start with some configuration, and we keep adding vertices and hyperedges over time as the process continues. The notion of preferential attachment is often informally described as the "rich get richer" phenomenon, meaning that the greater the degree of a vertex is, the greater the probability that it will be selected to a hyperedge to increase its degree. These two mechanisms were employed in the Barabási–Albert model [3] for random scale-free graphs and in the Avin et al. model [2] for random scale-free hypergraphs.

Another model, introduced by Fenner et al. [16], also uses these mechanisms. However, they consider another mechanism, namely vertex deactivation, to achieve a power-law distribution with an exponential cutoff. The idea of vertex deactivation is that during the process, vertices may also get deactivated with probability proportional to their degree (according to the preferential attachment mechanism). Such vertices are not removed from the graph, but they cannot be selected in the future, and thus, their degree freezes and never changes again. We can think of deactivation as if authors retire and stop publishing.

In this thesis, we generalize the Avin et al. model [2] by combining it with the Fenner et al. model [16] to generate random hypergraphs whose degree distribution follows a power-law distribution with an exponential cutoff. In Sections 3.1 and 3.2, we discuss notation and provide a formal definition of the model, and in Section 3.3, we analyse the degree distribution of random hypergraphs, generated according to the model.

3.1 Definitions and Notation

We define a *hypergraph* as a pair $H = (V, E)$, where V is a set of *vertices* and E is a multiset of non-empty sub-multisets of V, called *hyperedges*. We allow a hyperedge to appear multiple times in E since the same set of authors may publish several publications together. We also allow a vertex to appear in the same hyperedge multiple times in order to simplify further analysis of the model; however, as the model evolves, with high probability each vertex will be selected no more than once, given that sizes of hyperedges are appropriately small.

We define the degree of a vertex $v \in V$ in a hyperedge $e \in E$, denoted $\deg_e v$, to be the multiplicity of v in e, that is, the number of times v appears in e. The degree of the vertex v in the hypergraph is defined as $\deg v = \sum_{e \in E} \deg_e v$. The degree distribution $P(k)$ of a hypergraph H is the fraction of vertices of degree k, or formally,

$$P(k) = \frac{|\{v \in V \mid \deg v = k\}|}{|V|}. \tag{3.1}$$

Finally, the preferential attachment mechanism states that the probability to select a vertex v from some subset of vertices $A \subseteq V$ of a hypergraph H is proportional to the degree of v in H. Formally,

$$p_A(v) = \frac{\deg v}{\sum_{u \in A} \deg u}. \tag{3.2}$$

Then, when we say "a vertex is selected preferentially", we mean that a vertex is selected randomly according to the preferential attachment mechanism.

3.2 Model Description

The model can be described with a 5-tuple of parameters $H(H_0, p_v, p_e, p_d, Y)$, where

- $H_0 = (V_0, E_0)$ represents the initial hypergraph. For simplicity, we are going to assume that $|V_0| = 1$ and $|E_0| = 1$. However, we only require that it contains at least a single vertex $v \in V_0$ with $\deg v \geq 1$, as it can be shown that the initial configuration does not affect the asymptotic behaviour of the model.
- p_v, p_e and p_d are probabilities of different events that occur during the process. Naturally, $p_v + p_e + p_d = 1$ must hold.
- $Y = (Y_1, Y_2, \dots)$ is a sequence of independent random variables, where Y_t represents the size of a hyperedge that is added in step $t \geq 1$.

Define a random process which lets the hypergraph grow over time.

- **Step** $t = 0$**.** Initialize the process with H_0 and consider its vertices *active*.
- **Step** $t > 0$**.** Form H_t from H_{t-1} by performing one of the following actions:
  - (a) with probability p_v, add a new active vertex v and also preferentially select a hyperedge e of size $Y_t - 1$ among active vertices of H_{t-1}, and then add $\{v\} \cup e$ to H_t;
  - (b) with probability p_e, preferentially select a hyperedge of size Y_t among active vertices of H_{t-1} and add it to H_t;

(c) with probability p_d, preferentially select a single active vertex to *deactivate* it; deactivated vertices remain in the hypergraph, but they cannot be selected during the following steps of the process.

Whenever we need to select a hyperedge of size Y_t in step $t \geq 1$, we perform Y_t independent selections of vertices in V_{t-1} using the preferential attachment mechanism. Note that the same vertex may be selected several times to the same hyperedge. However, the probability of that happening approaches 0 as $t \to \infty$, when Y_t is selected to be appropriately small.

Note that since, on average, we need to add vertices more often than deactivate them, it is naturally required for $p_v > p_d$ to hold [16]. Otherwise, the model will eventually stop working as soon as we run out of active vertices. Nonetheless, it is possible that the model finishes with this error even when the requirement holds. We consider such cases invalid, and we restart the model whenever such an event occurs.

Remark 1. The model generalizes Avin et al. model [2] by introducing the deactivation of vertices, which leads to an appearance of an exponential cutoff. Hence, by specifying $p_d = 0$, we remove the vertex deactivation mechanism completely and thus obtain the Avin et al. model, which generates scale-free hypergraphs.

Remark 2. The model can also be seen as a generalization of Fenner et al. model [16], which is described in terms of balls with attached pins. In this case, balls correspond to vertices and pins correspond to hyperedges of size 1. Hence, by specifying $Y_t = 1$, in each step of the model we can either add a ball with a single pin, or add a pin to an existing active ball, or deactivate an existing active ball. This is exactly the definition of the Fenner et al. model.

3.3 Degree Distribution Analysis

In this section, we analyse the degree distribution of hypergraphs, generated according to the proposed model, and under a few assumptions, prove that it follows a power-law distribution with an exponential cutoff.

3.3.1 Preliminaries

Let $A_{k,t}$ denote the number of active vertices and $I_{k,t}$ denote the number of inactive vertices of degree k at step $t \geq 0$. Then, the total number of vertices of degree k at step t can be expressed as $N_{k,t} = A_{k,t} + I_{k,t}$. Let N_t denote the total number of vertices and let also A_t denote the number of active vertices at step $t \geq 0$; that is,

$$A_t = \sum_{k=1}^{\infty} A_{k,t}. \tag{3.3}$$

Since we start with a single vertex and then, in each step, we add no more than 1 vertex, we have

$$A_t \leq N_t \leq 1 + t. \tag{3.4}$$

We now derive the expectation of the total number of vertices N_t and the number of active vertices A_t at step $t \geq 0$. First, observe that since in each step we add a vertex with probability p_v and vertices are never removed from the hypergraph, we have

$$\mathbb{E}[N_t] = 1 + p_v t. \tag{3.5}$$

Likewise, since we start with a single vertex and then, in each time step, we either add a new active vertex with probability p_v, or deactivate an active vertex with probability p_d, we derive that the expectation of A_t can be expressed as

$$\mathbb{E}[A_t] = 1 + (p_v - p_d)t. \tag{3.6}$$

Let D_t denote the total sum of degrees of active vertices at step $t \geq 0$; that is,

$$D_t = \sum_{k=1}^{\infty} k A_{k,t}. \tag{3.7}$$

Then, considering that with probability $p_v + p_e$ we add a hyperedge of size Y_t, and with probability p_d we deactivate an active vertex, we now express the expectation of D_t as

$$\mathbb{E}[D_t] = 1 + (p_v + p_e) \sum_{\tau=1}^{t} \mathbb{E}[Y_t] - p_d \sum_{\tau=1}^{t} \mathbb{E}[\Theta_\tau], \tag{3.8}$$

where Θ_τ is a random variable that represents the degree of a vertex chosen for deactivation in step $\tau \geq 1$.

Before we move to the main theorem about the degree distribution of generated hypergraphs, we need to define several assumptions about certain distributions associated with the model. We then also state two technical lemmas, which we will use later in the proof.

Assumption 1. $\mathbb{E}[Y_t] = \mu \in \mathbb{R}_{>0}$, for all $t > 0$.

Assumption 2. $\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbb{E}[\Theta_\tau] = \theta \in \mathbb{R}_{>0}$.

Assumption 3. $\mathbb{E}\left[\dfrac{A_{k,t}}{D_t}\right] = \dfrac{\mathbb{E}[A_{k,t}]}{\mathbb{E}[D_t]} + o(1)$.

Assumption 4. $\mathbb{E}\left[\dfrac{Y_t^2}{D_{t-1}^2}\right] = o\left(\dfrac{1}{t}\right)$.

These assumptions are (unfortunately) required to prove the main theorem of this thesis, and they also arise in [2] and [16]. Essentially, Assumptions 1 and 4 put restrictions on sizes of hyperedge, as in reality they are relatively small. Assumption 2 states that the average degree of a deactivated vertex converges to some constant. Note that from Assumptions 1 and 2, using Equation 3.8, we obtain

$$\lim_{t\to\infty} \frac{\mathbb{E}[D_t]}{t} = (p_v + p_e)\mu - p_d\theta, \tag{3.9}$$

and also, considering that $\mathbb{E}[\Theta_\tau] \geq 1$ and $\mathbb{E}[A_t] \leq \mathbb{E}[D_t]$, we have

$$1 \leq \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}[\Theta_\tau] \leq 1 + \frac{p_v(\mu - 1) + p_e\mu}{p_d}. \tag{3.10}$$

Finally, note that Assumption 3 is a general technical assumption and is only required to prove the main theorem. It could be replaced with a more "natural" assumption about the concentration of D_t around its expectation (if we assume that with high probability $D_t = \mathbb{E}[D_t] + o(t)$). In Chapter 4, we provide convincing simulation results to support Assumptions 2 and 3.

Lemma 1 ([8, Chapter 3.3]). *Let (a_t), (b_t), (c_t) be three sequences such that*

$$a_{t+1} = a_t\left(1 - \frac{b_t}{t}\right) + c_t, \text{ for any } t \geq t_0, \tag{3.11}$$

$\lim_{t\to\infty} b_t = b > 0$ and $\lim_{t\to\infty} c_t = c$. Then $\lim_{t\to\infty} \dfrac{a_t}{t}$ exists and equals $\dfrac{c}{1+b}$.

Lemma 2 ([18, Lemma 4]). *Let $N_{k,t}$ denote the number of vertices of degree k and N_t denote the number of vertices at step $t \geq 0$. Then*

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{N_{k,t}}{N_t}\right] = \lim_{t\to\infty} \frac{\mathbb{E}[N_{k,t}]}{\mathbb{E}[N_t]}. \tag{3.12}$$

3.3.2 Main Theorem

Having introduced the notation and assumptions, we proceed to the main theorem about the degree distribution of hypergraphs generated according to the proposed model.

Theorem 3. *Let H be a hypergraph generated according to the model $H(H_0, p_v, p_e, p_d, Y)$, and suppose that Assumptions 1–4 hold. Then the degree distribution $P(k)$ of H follows a power-law distribution with an exponential cutoff of the form*

$$P(k) \sim \frac{C}{p_v} \frac{\gamma^k}{k^{1/\beta}} \left(\frac{1}{k} + \delta \right), \tag{3.13}$$

where constants C, β, γ and δ can be expressed in terms of parameters of the model.

Proof. We closely follow the master equation approach from [2]. The proof is organized as follows: we first derive and solve a recurrence equation for the expected number of active vertices of degree k, which then allows us to derive and solve a similar recurrence equation for the expected number of inactive vertices of degree k. Finally, we use these two results to evaluate the degree distribution.

Evaluating $\mathbb{E}[A_{k,t}]/t$ as $t \to \infty$ Remind that $A_{k,t}$ denotes the number of active vertices of degree k at step t. To evaluate $\lim_{t\to\infty} \mathbb{E}[A_{k,t}]/t$, we use mathematical induction on k. We first consider the case $k = 1$. An active vertex remains in $A_{1,t}$ if it had degree 1 at step $t-1$ and was neither selected to a hyperedge, nor deactivated. Also, in each step, with probability p_v, a single new active vertex of degree 1 is added to the hypergraph. Now, let $\mathcal{F}_t$ denote a σ-algebra associated with the probability space at step t. Then

$$\begin{aligned} \mathbb{E}[A_{1,t} \,|\, \mathcal{F}_{t-1}] = p_v A_{1,t-1} \left(1 - \frac{1}{D_{t-1}} \right)^{Y_t - 1} + p_e A_{1,t-1} \left(1 - \frac{1}{D_{t-1}} \right)^{Y_t} \\ + p_d A_{1,t-1} \left(1 - \frac{1}{D_{t-1}} \right) + p_v. \end{aligned} \tag{3.14}$$

We now derive bounds for $\mathbb{E}[A_{1,t}]$. By taking the expectation of both sides of Equation 3.14 and by using Bernoulli's inequality, we obtain

$$\begin{aligned} \mathbb{E}[A_{1,t}] &\geq p_v \mathbb{E}\left[A_{1,t-1} \left(1 - \frac{Y_t - 1}{D_{t-1}} \right) \right] + p_e \mathbb{E}\left[A_{1,t-1} \left(1 - \frac{Y_t}{D_{t-1}} \right) \right] \\ &\quad + p_d \mathbb{E}\left[A_{1,t-1} \left(1 - \frac{1}{D_{t-1}} \right) \right] + p_v \\ &= p_v \mathbb{E}[A_{1,t-1}] \left(1 - \frac{\mathbb{E}[Y_t] - 1}{\mathbb{E}[D_{t-1}]} \right) + p_e \mathbb{E}[A_{1,t-1}] \left(1 - \frac{\mathbb{E}[Y_t]}{\mathbb{E}[D_{t-1}]} \right) \\ &\quad + p_d \mathbb{E}[A_{1,t-1}] \left(1 - \frac{1}{\mathbb{E}[D_{t-1}]} \right) + p_v - o(1) \\ &= \mathbb{E}[A_{1,t-1}] \left(1 - \frac{p_v(\mu - 1) + p_e \mu + p_d}{\mathbb{E}[D_{t-1}]} \right) + p_v - o(1), \end{aligned} \tag{3.15}$$

where the penultimate equality follows from the assumptions we have made about the distribution of $\mathbb{E}[A_{1,t-1}/D_{t-1}]$ and the independence of Y_t from $A_{1,t-1}$ and D_{t-1}. On the other hand, since $(1-x)^n \leq 1/(1+nx)$ for $x \in [0,1]$ and $n \in \mathbb{N}$, we have

$$\begin{aligned}
\mathbb{E}[A_{1,t}] &\leq p_v \mathbb{E}\left[\frac{A_{1,t-1}}{1+(Y_t-1)/D_{t-1}}\right] + p_e \mathbb{E}\left[\frac{A_{1,t-1}}{1+Y_t/D_{t-1}}\right] \\
&\quad + p_d \mathbb{E}\left[A_{1,t-1}\left(1-\frac{1}{D_{t-1}}\right)\right] + p_v \\
&= p_v \mathbb{E}\left[A_{1,t-1}\left(1-\frac{Y_t-1}{D_{t-1}+Y_t-1}\right)\right] + p_e \mathbb{E}\left[A_{1,t-1}\left(1-\frac{Y_t}{D_{t-1}+Y_t}\right)\right] \\
&\quad + p_d \mathbb{E}\left[A_{1,t-1}\left(1-\frac{1}{D_{t-1}}\right)\right] + p_v \\
&\leq p_v \mathbb{E}\left[A_{1,t-1}\left(1-\frac{Y_t-1}{D_{t-1}}+\frac{(Y_t-1)^2}{D_{t-1}^2}\right)\right] + p_e \mathbb{E}\left[A_{1,t-1}\left(1-\frac{Y_t}{D_{t-1}}+\frac{Y_t^2}{D_{t-1}^2}\right)\right] \\
&\quad + p_d \mathbb{E}\left[A_{1,t-1}\left(1-\frac{1}{D_{t-1}}\right)\right] + p_v \\
&= p_v \mathbb{E}[A_{1,t-1}]\left(1-\frac{\mathbb{E}[Y_t]-1}{\mathbb{E}[D_{t-1}]}\right) + p_e \mathbb{E}[A_{1,t-1}]\left(1-\frac{\mathbb{E}[Y_t]}{\mathbb{E}[D_{t-1}]}\right) \\
&\quad + p_d \mathbb{E}[A_{1,t-1}]\left(1-\frac{1}{\mathbb{E}[D_{t-1}]}\right) + p_v + p_v \mathbb{E}\left[O(t)\frac{(Y_t-1)^2}{D_{t-1}^2}\right] + p_e \mathbb{E}\left[O(t)\frac{Y_t^2}{D_{t-1}^2}\right] \\
&= \mathbb{E}[A_{1,t-1}]\left(1-\frac{p_v(\mu-1)+p_e\mu+p_d}{\mathbb{E}[D_{t-1}]}\right) + p_v + o(1).
\end{aligned} \tag{3.16}$$

From Equations 3.15 and 3.16, we thus conclude that

$$\mathbb{E}[A_{1,t}] = \mathbb{E}[A_{1,t-1}]\left(1-\frac{p_v(\mu-1)+p_e\mu+p_d}{\mathbb{E}[D_{t-1}]}\right) + p_v + o(1). \tag{3.17}$$

We now apply Lemma 1 to evaluate $\lim_{t\to\infty}\mathbb{E}[A_{1,t}]/t$. Let

$$a_t = \mathbb{E}[A_{1,t}], \tag{3.18}$$

$$b_t = \frac{p_v(\mu-1)+p_e\mu+p_d}{\mathbb{E}[D_{t-1}]/t}, \tag{3.19}$$

$$c_t = p_v + o(1). \tag{3.20}$$

From Equation 3.9, we then obtain that

$$\lim_{t\to\infty} b_t = \frac{p_v(\mu-1)+p_e\mu+p_d}{\mu(p_v+p_e)-p_d\theta} = \beta, \tag{3.21}$$

$$\lim_{t\to\infty} c_t = p_v, \tag{3.22}$$

and thus,

$$\lim_{t\to\infty}\frac{a_t}{t}=\lim_{t\to\infty}\frac{\mathbb{E}[A_{1,t}]}{t}=\bar{A}_1=\frac{p_v}{1+\beta}. \tag{3.23}$$

Now, assume the induction hypothesis that $\lim_{t\to\infty}\mathbb{E}[A_{k-1,t}]/t$ exists and equals $\bar{A}_{k-1}$. For $k>1$, an active vertex appears in $A_{k,t}$ if it was active at step $t-1$, had degree $k-l$ and was chosen exactly l times to a hyperedge, or it had degree k and was not selected for deactivation. Thus,

$$\begin{aligned}\mathbb{E}[A_{k,t}\,|\,\mathcal{F}_{t-1}]=p_v&\sum_{l=0}^{\min\{Y_t-1,k-1\}}A_{k-l,t-1}\mathbb{P}\left[l,Y_t-1,\frac{k-l}{D_{t-1}}\right]+p_e\sum_{l=0}^{\min\{Y_t,k-1\}}A_{k-l,t-1}\mathbb{P}\left[l,Y_t,\frac{k-l}{D_{t-1}}\right]\\&+p_dA_{k,t-1}\left(1-\frac{k}{D_{t-1}}\right),\end{aligned} \tag{3.24}$$

where

$$\mathbb{P}[l,n,p]=\binom{n}{l}p^l(1-p)^{n-l}. \tag{3.25}$$

We now derive bounds for $\mathbb{E}[A_{k,t}]$. Let

$$\mathbb{E}[A_{k,t}]=\mathbb{E}[\psi]+p_v\mathbb{E}[\varphi(Y_t-1)]+p_e\mathbb{E}[\varphi(Y_t)], \tag{3.26}$$

where

$$\begin{aligned}\psi&=p_v\sum_{l=0}^{1}A_{k-l,t-1}\mathbb{P}\left[l,Y_t-1,\frac{k-l}{D_{t-1}}\right]+p_e\sum_{l=0}^{1}A_{k-l,t-1}\mathbb{P}\left[l,Y_t,\frac{k-l}{D_{t-1}}\right]\\&\quad+p_dA_{k,t-1}\left(1-\frac{k}{D_{t-1}}\right)\\&=A_{k,t-1}\left(p_v\left(1-\frac{k}{D_{t-1}}\right)^{Y_t-1}+p_e\left(1-\frac{k}{D_{t-1}}\right)^{Y_t}+p_d\left(1-\frac{k}{D_{t-1}}\right)\right)\\&\quad+A_{k-1,t-1}\frac{k-1}{D_{t-1}}\left(p_v(Y_t-1)\left(1-\frac{k-1}{D_{t-1}}\right)^{Y_t-2}+p_eY_t\left(1-\frac{k-1}{D_{t-1}}\right)^{Y_t-1}\right)\end{aligned} \tag{3.27}$$

and

$$\varphi(n)=\sum_{l=2}^{\min\{n,k-1\}}A_{k-l,t-1}\mathbb{P}\left[l,n,\frac{k-l}{D_{t-1}}\right]. \tag{3.28}$$

We aim to show that only the term $\mathbb{E}[\psi]$ is significant and that the terms $\mathbb{E}[\varphi(Y_t - 1)]$ and $\mathbb{E}[\varphi(Y_t)]$ converge to 0 as $t \to \infty$. We have

$$\begin{aligned}
\varphi(Y_t) &\le \sum_{l=2}^{k-1} A_{k-l,t-1} \binom{Y_t}{l} \left(\frac{k-l}{D_{t-1}}\right)^l \left(1 - \frac{k-l}{D_{t-1}}\right)^{Y_t - l} \\
&\le O(t) \sum_{l=2}^{k-1} \binom{Y_t}{l} \left(\frac{k-l}{D_{t-1}}\right)^l \left(1 - \frac{k-l}{D_{t-1}}\right)^{Y_t - l} \\
&\le O(t) \sum_{l=2}^{k-1} Y_t^l \left(\frac{k}{D_{t-1}}\right)^l \left(1 - \frac{1}{D_{t-1}}\right)^{Y_t - k + 1} \\
&\le O(t) \frac{Y_t^2 k^2}{D_{t-1}^2} e^{-Y_t/D_{t-1}} e^{k-1} \sum_{l=2}^{k-1} \left(\frac{Y_t k}{D_{t-1}}\right)^{l-2} \\
&= O(t) \frac{Y_t^2}{D_{t-1}^2} e^{-Y_t/D_{t-1}} \sum_{l=2}^{k-1} \left(\frac{Y_t k}{D_{t-1}}\right)^{l-2}.
\end{aligned} \tag{3.29}$$

Then, if $Y_t \le D_{t-1}$, we have

$$\varphi(Y_t) \le O(t) \frac{Y_t^2}{D_{t-1}^2} k^{k-2} = O(t) \frac{Y_t^2}{D_{t-1}^2}. \tag{3.30}$$

Otherwise, we have

$$\begin{aligned}
\varphi(Y_t) &\le O(t) \frac{Y_t^2}{D_{t-1}^2} e^{-Y_t/D_{t-1}} \frac{(Y_t k / D_{t-1})^{k-2} - 1}{(Y_t k / D_{t-1}) - 1} \\
&\le O(t) \frac{Y_t^2}{D_{t-1}^2} e^{-Y_t/D_{t-1}} \frac{(Y_t / D_{t-1})^{k-2}}{k-1} k^{k-2} \\
&\le O(t) \frac{Y_t^2}{D_{t-1}^2} e^{-(k-2)} \frac{(k-2)^{k-2}}{k-1} k^{k-2} \\
&= O(t) \frac{Y_t^2}{D_{t-1}^2}.
\end{aligned} \tag{3.31}$$

where the last inequality follows from the fact that $e^{-x} x^{\alpha}$ is maximized at $x = \alpha$. Therefore, under the assumptions we have made about the distributions of Y_t and D_{t-1}, in both cases we have $\mathbb{E}[\varphi(Y_t)] = o(1)$ and similarly, $\mathbb{E}[\varphi(Y_t - 1)] = o(1)$. We now derive bounds for $\mathbb{E}[\psi]$, similarly to what we have done for $\mathbb{E}[A_{1,t}]$. We thus have

$$\mathbb{E}[\psi] = \mathbb{E}\left[A_{k,t-1}\left(p_v\left(1 - \frac{k}{D_{t-1}}\right)^{Y_t - 1} + p_e\left(1 - \frac{k}{D_{t-1}}\right)^{Y_t} + p_d\left(1 - \frac{k}{D_{t-1}}\right)\right)\right]$$

$$
\begin{aligned}
&+ \mathbb{E}\left[A_{k-1,t-1}\frac{k-1}{D_{t-1}}\left(p_v(Y_t-1)\left(1-\frac{k-1}{D_{t-1}}\right)^{Y_t-2} + p_eY_t\left(1-\frac{k-1}{D_{t-1}}\right)^{Y_t-1}\right)\right] \\
\geq\ & \mathbb{E}\left[A_{k,t-1}\left(p_v\left(1-\frac{(Y_t-1)k}{D_{t-1}}\right) + p_e\left(1-\frac{Y_tk}{D_{t-1}}\right) + p_d\left(1-\frac{k}{D_{t-1}}\right)\right)\right] \\
&+ \mathbb{E}\left[A_{k-1,t-1}\frac{k-1}{D_{t-1}}\left(1-\frac{(k-1)(Y_t-1)}{D_{t-1}}\right)(p_v(Y_t-1)+p_eY_t)\right] \\
=\ & \mathbb{E}[A_{k,t-1}]\left(1-\frac{k(p_v(\mu-1)+p_e\mu+p_d)}{\mathbb{E}[D_{t-1}]}\right) \\
&+ \mathbb{E}[A_{k-1,t-1}]\frac{(k-1)(p_v(\mu-1)+p_e\mu)}{\mathbb{E}[D_{t-1}]} + o(1). \qquad (3.32)
\end{aligned}
$$

At the same time,

$$
\begin{aligned}
\mathbb{E}[\psi] &\leq \mathbb{E}[A_{k,t-1}]\left(1-\frac{k(p_v(Y_t-1)+p_eY_t+p_d)}{D_{t-1}}\right) + \mathbb{E}\left[O(t)\frac{Y_t^2}{D_{t-1}^2}\right] \\
&+ \mathbb{E}\left[A_{k-1,t-1}\frac{k-1}{D_{t-1}}(p_v(Y_t-1)+p_eY_t)\right] \\
&= \mathbb{E}[A_{k,t-1}]\left(1-\frac{k(p_v(\mu-1)+p_e\mu+p_d)}{\mathbb{E}[D_{t-1}]}\right) \\
&+ \mathbb{E}[A_{k-1,t-1}]\frac{(k-1)(p_v(\mu-1)+p_e\mu)}{\mathbb{E}[D_{t-1}]} + o(1). \qquad (3.33)
\end{aligned}
$$

From Equations 3.32 and 3.33, we hence obtain

$$
\begin{aligned}
\mathbb{E}[A_{k,t}] = \mathbb{E}[A_{k,t-1}]&\left(1-\frac{k(p_v(\mu-1)+p_e\mu+p_d)}{\mathbb{E}[D_{t-1}]}\right) \\
&+ \mathbb{E}[A_{k-1,t-1}]\frac{(k-1)(p_v(\mu-1)+p_e\mu)}{\mathbb{E}[D_{t-1}]} + o(1).
\end{aligned} \qquad (3.34)
$$

Remind that by the induction hypothesis, we assume that $\lim_{t\to\infty}\mathbb{E}[A_{k-1,t}]/t = \bar{A}_{k-1}$. We apply Lemma 1 to evaluate $\lim_{t\to\infty}\mathbb{E}[A_{k,t}]/t$. Let

$$a_t = \mathbb{E}[A_{k,t}], \qquad (3.35)$$

$$b_t = \frac{k(p_v(\mu-1)+p_e\mu+p_d)}{\mathbb{E}[D_{t-1}]/t}, \qquad (3.36)$$

$$c_t = \frac{\mathbb{E}[A_{k-1,t-1}]}{t}\frac{(k-1)(p_v(\mu-1)+p_e\mu)}{\mathbb{E}[D_{t-1}]/t} + o(1). \qquad (3.37)$$

We therefore obtain that

$$\lim_{t\to\infty} b_t = k\beta \qquad (3.38)$$

and

$$\lim_{t\to\infty} c_t = \bar{A}_{k-1}\frac{(k-1)(p_v(\mu-1)+p_e\mu)}{\mu(p_v+p_e)-p_d\theta}. \tag{3.39}$$

Hence,

$$\lim_{t\to\infty}\frac{a_t}{t} = \lim_{t\to\infty}\frac{\mathbb{E}[A_{k,t}]}{t} = \bar{A}_k = \bar{A}_{k-1}\frac{(k-1)\gamma}{k+1/\beta}, \tag{3.40}$$

where

$$\gamma = \frac{p_v(\mu-1)+p_e\mu}{p_v(\mu-1)+p_e\mu+p_d}. \tag{3.41}$$

We finally have

$$\bar{A}_1 = \frac{p_v}{\beta}\frac{1}{(1+1/\beta)}, \tag{3.42}$$

$$\bar{A}_2 = \frac{p_v}{\beta}\frac{\gamma}{(1+1/\beta)(2+1/\beta)}, \tag{3.43}$$

$$\dots$$

$$\bar{A}_k = \frac{p_v}{\beta}\frac{\gamma^{k-1}(k-1)!}{(1+1/\beta)(2+1/\beta)\dots(k+1/\beta)}. \tag{3.44}$$

Hence,

$$\bar{A}_k = \frac{p_v}{\beta\gamma}\frac{\gamma^k\Gamma(1+1/\beta)\Gamma(k)}{\Gamma(k+1+1/\beta)} \sim C\frac{\gamma^k}{k^{1+1/\beta}}, \tag{3.45}$$

where

$$C = \frac{p_v}{\beta\gamma}\Gamma(1+1/\beta). \tag{3.46}$$

Evaluating $\mathbb{E}[I_{k,t}]/t$ as $t \to \infty$ We now observe that the expected number of inactive vertices of degree $k \geq 1$ at step t, given $\mathcal{F}_{t-1}$, can be expressed as

$$\mathbb{E}[I_{k,t} \,|\, \mathcal{F}_{t-1}] = I_{k,t-1} + p_d A_{k,t-1}\frac{k}{D_{t-1}} \tag{3.47}$$

since inactive vertices of degree k remain in $I_{k,t}$ forever and a vertex of degree k becomes inactive if it was selected in step $t-1$ for deactivation. By taking the expectation of both sides, we obtain

$$\mathbb{E}[I_{k,t}] = \mathbb{E}[I_{k,t-1}] + p_d\mathbb{E}[A_{k,t-1}]\frac{k}{\mathbb{E}[D_{t-1}]} + o(1). \tag{3.48}$$

We then have that

$$\lim_{t\to\infty}(\mathbb{E}[I_{k,t}] - \mathbb{E}[I_{k,t-1}]) = \lim_{t\to\infty} p_d k\frac{\mathbb{E}[A_{k,t-1}]}{t}\frac{t}{\mathbb{E}[D_{t-1}]} + o(1)$$

$$
\begin{aligned}
&= \bar{A}_k \frac{p_d k}{(p_v + p_e)\mu - p_d\theta} \\
&= \bar{A}_k k\delta, \qquad (3.49)
\end{aligned}
$$

where

$$\delta = \frac{p_d}{(p_v + p_e)\mu - p_d\theta}. \qquad (3.50)$$

Hence, by Stolz–Cesàro theorem, we obtain

$$\bar{I}_k = \lim_{t\to\infty} \frac{\mathbb{E}[I_{k,t}]}{t} = \lim_{t\to\infty} (\mathbb{E}[I_{k,t}] - \mathbb{E}[I_{k,t-1}]) = \bar{A}_k k\delta. \qquad (3.51)$$

Degree Distribution Remind that $N_{k,t}$ denotes the total number of vertices at step t; that is, $N_{k,t} = A_{k,t} + I_{k,t}$. Therefore,

$$\bar{N}_k = \lim_{t\to\infty} \frac{\mathbb{E}[N_{k,t}]}{t} = \bar{A}_k + \bar{I}_k = \bar{A}_k(1 + k\delta) \sim C\frac{\gamma^k}{k^{1/\beta}}\left(\frac{1}{k} + \delta\right). \qquad (3.52)$$

To refer to the degree distribution of a hypergraph, we need to consider the limit of the expected fraction of vertices of degree k as $t \to \infty$. Hence, by applying Lemma 2, we obtain

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{N_{k,t}}{N_t}\right] = \lim_{t\to\infty} \frac{\mathbb{E}[N_{k,t}]}{\mathbb{E}[N_t]} = \frac{\bar{N}_k}{p_v} \sim \frac{C}{p_v}\frac{\gamma^k}{k^{1/\beta}}\left(\frac{1}{k} + \delta\right), \qquad (3.53)$$

which thereby concludes the proof. □

3.3.3 Evaluating θ

In Theorem 3, we have proved that the degree distribution of a hypergraph, generated according to the model $H(H_0, p_v, p_e, p_d, Y)$, follows a power-law distribution with an exponential cutoff. The parameters of this theoretical distribution can be expressed in terms of p_v, p_e, p_d, μ and θ. Recall that Θ_t is a random variable that describes the degree of a vertex selected for deactivation in step $t \geq 1$. Also, from Assumption 2, remind that

$$\theta = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbb{E}[\Theta_\tau]. \qquad (3.54)$$

However, we do not know how Θ_t is distributed, and thus, we cannot express θ. In this section, we discuss a method, which, under a few assumptions, allows us to evaluate θ with arbitrary precision.

From the definition of Θ_t, we can express

$$\mathbb{E}[\Theta_t] = \mathbb{E}\left[\frac{\sum_{k=1}^{\infty} k^2 A_{k,t}}{\sum_{k=1}^{\infty} k A_{k,t}}\right]. \tag{3.55}$$

We now observe that from the expression of $\bar{A}_k$, we can obtain

$$\sum_{k=1}^{\infty} k\bar{A}_k = \frac{p_v}{1+\beta}\sum_{n=0}^{\infty}\frac{\Gamma(n+1)\Gamma(n+2)\Gamma(2+1/\beta)}{\Gamma(n+2+1/\beta)}\frac{\gamma^n}{n!} = \frac{p_v}{1+\beta}\sum_{n=0}^{\infty}\frac{(1)_n(2)_n}{(2+1/\beta)_n}\frac{\gamma^n}{n!}, \tag{3.56}$$

where $(x)_n = \Gamma(x+n)/\Gamma(x)$ is the Pochhammer symbol. We can further nicely express this series using the Gaussian hypergeometric function [1]

$$F(a,b;c;z) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!}, \tag{3.57}$$

which is defined for $|z| < 1$ and $c \notin \mathbb{Z}_{\leq 0}$. Thus,

$$\sum_{k=1}^{\infty} k\bar{A}_k = \frac{p_v}{1+\beta}F(1,2;2+1/\beta;\gamma), \tag{3.58}$$

and similarly,

$$\sum_{k=1}^{\infty} k^2\bar{A}_k = \frac{p_v}{1+\beta}F(2,2;2+1/\beta;\gamma). \tag{3.59}$$

If we now assume that $\mathbb{E}[\Theta_t]$ can be approximated by

$$\mathbb{E}[\Theta_t] = \frac{\mathbb{E}[\sum_{k=1}^{\infty} k^2 A_{k,t}]}{\mathbb{E}[\sum_{k=1}^{\infty} k A_{k,t}]} + o(1), \tag{3.60}$$

then, using Equations 3.58 and 3.59, we obtain

$$\lim_{t\to\infty}\mathbb{E}[\Theta_t] = \lim_{t\to\infty}\frac{\mathbb{E}[\sum_{k=1}^{\infty} k^2 A_{k,t}]}{\mathbb{E}[\sum_{k=1}^{\infty} k A_{k,t}]} + o(1) = \frac{\sum_{k=1}^{\infty} k^2\bar{A}_k}{\sum_{k=1}^{\infty} k\bar{A}_k} = \frac{F(2,2;2+1/\beta;\gamma)}{F(1,2;2+1/\beta;\gamma)}. \tag{3.61}$$

We can then apply Stolz–Cesàro theorem to obtain

$$\theta = \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{t}\mathbb{E}[\Theta_\tau] = \lim_{t\to\infty}\mathbb{E}[\Theta_t] = \frac{F(2,2;2+1/\beta;\gamma)}{F(1,2;2+1/\beta;\gamma)}. \tag{3.62}$$

Remind that

$$\frac{1}{\beta} = \frac{(p_v+p_e)\mu - p_d\theta}{p_v(\mu-1)+p_e\mu+p_d}, \tag{3.63}$$

and let $F_1(\theta) = F(1, 2; \rho(\theta); \gamma)$ and $F_2(\theta) = F(2, 2; \rho(\theta); \gamma)$, where

$$\rho(\theta) = 2 + \frac{(p_v + p_e)\mu - p_d\theta}{p_v(\mu - 1) + p_e\mu + p_d}. \tag{3.64}$$

We now define

$$R(\theta) = \frac{F_2(\theta)}{F_1(\theta)}. \tag{3.65}$$

We have thus obtained that θ is a fixed point of the function $R(\theta)$. One of the possible ways to find a fixed point of a function is to apply the fixed-point iteration method, which consists of calculating the limit of the sequence

$$\theta_{n+1} = R(\theta_n), \tag{3.66}$$

starting from some initial value θ_0. However, there are several things we need to consider before we can apply this method. First, not every such sequence has a limit, and it may be the case that it will diverge when starting from improper θ_0. Second, there may be several fixed points of $R(\theta)$, thus leading to ambiguity. We, therefore, proceed to analyse the behaviour of $R(\theta)$ to show that it always has a *unique* fixed point, which can always be found using the fixed-point iteration method.

Lemma 4.

$$R(\theta) = \theta - \frac{p_v}{p_d} + \frac{1}{1-\gamma}\frac{\rho(\theta) - 1}{F_1(\theta)}. \tag{3.67}$$

Proof. We can use the Gauss' contiguous relations [1], namely

$$a(F(a+) - F) = \frac{(c-a)F(a-) + (a - c + bz)F}{1 - z}, \tag{3.68}$$

where $F = F(a, b; c; z)$, $F(a+) = F(a + 1, b; c; z)$ and $F(a-) = F(a - 1, b; c; z)$. From Equation 3.68, we obtain

$$F(a+) = \frac{\big(2a - c + (b - a)z\big)F + (c - a)F(a-)}{a(1 - z)}, \tag{3.69}$$

and thus

$$\frac{F(a+)}{F} = \frac{2a - c + (b - a)z}{a(1 - z)} + \frac{(c - a)F(a-)}{a(1 - z)F}. \tag{3.70}$$

We observe that by definition of the hypergeometric function, $F(0, b; c; z) = 1$. By plugging $a = 1$, $b = 2$, $c = \rho(\theta)$ and $z = \gamma$ into Equation 3.70, we obtain Equation 3.67. $\square$

We further proceed to analyse the behaviour of $R(\theta)$ on the interval $[0, \hat{\theta}]$, where

$$\hat{\theta} = \frac{(p_v + p_e)\mu}{p_d}. \tag{3.71}$$

There are several reasons for selecting such $\hat{\theta}$:

1. From Equation 3.10, remind that

$$1 \leq \theta \leq \frac{p_v(\mu - 1) + p_e\mu + p_d}{p_d}. \tag{3.72}$$

However, we require $p_v > p_d$, which implies that $\theta \leq \hat{\theta}$. Hence, the fixed point of $R(\theta)$ that we are looking for must definitely belong to the interval $[0, \hat{\theta}]$.

2. $\rho(\hat{\theta}) = 2$, which makes it very easy to evaluate the hypergeometric function at that point. For example, $F_1(\hat{\theta})$ is a simple geometric series with the ratio γ. Then, using Equation 3.67, we can show that

$$R(\hat{\theta}) = \hat{\theta} - \frac{p_v}{p_d} + 1. \tag{3.73}$$

3. Remind that the hypergeometric function $F(a, b; c; z)$ is defined for $|z| < 1$ and $c \notin \mathbb{Z}_{\leq 0}$. Therefore, since $0 < \gamma < 1$ and $\rho(\theta)$ is positive on $[0, \hat{\theta}]$, both $F_1(\theta)$ and $F_2(\theta)$ are always defined and positive on $[0, \hat{\theta}]$.
4. $R(\theta)$ is continuous on $[0, \hat{\theta}]$ since both $F_1(\theta)$ and $F_2(\theta)$ are continuous and $F_1(\theta)$ is positive on $[0, \hat{\theta}]$.
5. We know that $R(0) > 0$ and $R(\hat{\theta}) < \hat{\theta}$. Then, since $R(\theta)$ is continuous on $[0, \hat{\theta}]$, by the mean value theorem, there exists at least a single fixed point within the specified interval.

Lemma 5. *$R(\theta)$ strictly increases on* $[0, \hat{\theta}]$.

Proof. First, note that the derivative of the hypergeometric function with respect to c is

$$F'(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!} \big(\psi(c) - \psi(c + n)\big), \tag{3.74}$$

where $\psi(x)$ denotes the digamma function. Since the digamma function increases on $(0, +\infty)$, we can see that $F'(a, b; c; z)$ is negative when parameters of the function are positive.

Now, observe that since $\rho'(\theta) = \gamma - 1$, we have

$$F_1'(\theta) = (\gamma - 1)F'(1, 2; \rho(\theta); \gamma) \quad \text{and} \quad F_2'(\theta) = (\gamma - 1)F'(2, 2; \rho(\theta); \gamma), \tag{3.75}$$

and thus, given that $\gamma < 1$, they are both positive. In order to determine the sign of

$$R'(\theta) = \frac{F_2'(\theta)F_1(\theta) - F_2(\theta)F_1'(\theta)}{F_1(\theta)^2}, \tag{3.76}$$

we only need to determine the sign of its numerator. By considering the Cauchy product of $F_2'(\theta)$ and $F_1(\theta)$, we obtain

$$F_2'(\theta)F_1(\theta) = (\gamma - 1)\sum_{k=0}^{\infty}\gamma^k\sum_{n=0}^{k}(n+1)\frac{(2)_n}{(\rho(\theta))_n}\frac{(2)_{k-n}}{(\rho(\theta))_{k-n}}\big(\psi(\rho(\theta)) - \psi(\rho(\theta)+n)\big). \tag{3.77}$$

Similarly, for $F_2(\theta)$ and $F_1'(\theta)$, we have

$$F_2(\theta)F_1'(\theta) = (\gamma - 1)\sum_{k=0}^{\infty}\gamma^k\sum_{n=0}^{k}(k-n+1)\frac{(2)_n}{(\rho(\theta))_n}\frac{(2)_{k-n}}{(\rho(\theta))_{k-n}}\big(\psi(\rho(\theta)) - \psi(\rho(\theta)+n)\big). \tag{3.78}$$

Finally, we express the difference between these two expressions as

$$\begin{aligned} &F_2'(\theta)F_1(\theta) - F_2(\theta)F_1'(\theta) = \\ &\qquad (\gamma - 1)\sum_{k=0}^{\infty}\gamma^k\sum_{n=0}^{k}(2n-k)\frac{(2)_n}{(\rho(\theta))_n}\frac{(2)_{k-n}}{(\rho(\theta))_{k-n}}\big(\psi(\rho(\theta)) - \psi(\rho(\theta)+n)\big). \end{aligned} \tag{3.79}$$

We now check the sign of the inner sum. Observe that the sum of two elements with indices n and $k-n$ is

$$(2n-k)\frac{(2)_n}{(\rho(\theta))_n}\frac{(2)_{k-n}}{(\rho(\theta))_{k-n}}\big(\psi(\rho(\theta)+k-n) - \psi(\rho(\theta)+n)\big). \tag{3.80}$$

Since the digamma function increases on $(0, +\infty)$, we conclude that the inner sum is negative, which implies $F_2'(\theta)F_1(\theta) - F_2(\theta)F_1'(\theta) > 0$ and concludes that $R'(\theta) > 0$. □

Lemma 6. *$R(\theta)$ is a contraction mapping on $[0, \hat{\theta}]$.*

Proof. Remind that a function $f : X \mapsto X$, defined on a metric space (X, d), is called a contraction mapping, if there exists a constant $q \in [0, 1)$, such that for all $x_1, x_2 \in X$, we have

$$d(f(x_1), f(x_2)) \leq qd(x_1, x_2). \tag{3.81}$$

If $f(x)$ is a differentiable function, such that $\sup|f'(x)| < 1$, then $f(x)$ is a contraction mapping with $q = \sup|f'(x)|$.

Now, from Equation 3.67, we obtain

$$R'(\theta) = 1 + \frac{1}{1-\gamma} \frac{\rho'(\theta)F_1(\theta) - (\rho(\theta)-1)F_1'(\theta)}{F_1(\theta)^2}. \tag{3.82}$$

Remind that $F_1(\theta)$ is positive and increases, and $\rho(\theta) > 1$ and decreases on $[0, \hat{\theta}]$, which implies that the right term of the expression is negative. Since $R(\theta)$ also increases on $[0, \hat{\theta}]$, we have that $|R'(\theta)| \in [0, 1)$ for any $\theta \in [0, \hat{\theta}]$. Therefore, by the extreme value theorem, we know that $|R'(\theta)|$ achieves some maximum value $q \in (0, 1)$. Then, since $[0, \hat{\theta}]$ is a complete metric space and $R([0, \hat{\theta}]) \subseteq [0, \hat{\theta}]$, we conclude that $R(\theta)$ is a contraction mapping on $[0, \hat{\theta}]$. □

Theorem 7. *$R(\theta)$ has a unique fixed point θ^* in $[0, \hat{\theta}]$, such that*

$$\lim_{n\to\infty} \theta_n = \theta^*, \tag{3.83}$$

where $\theta_{n+1} = R(\theta_n)$ and θ_0 can take any value in $[0, \hat{\theta}]$.

Proof. The proof directly follows from the fact that $R(\theta)$ is a contraction mapping, defined on a complete metric space, and the Banach fixed-point theorem. □

Remark. We have thus shown that the fixed-point iteration method can be used to find θ as the fixed point θ^* of the function $R(\theta)$. In fact, we can use the constant q to describe the speed of convergence of θ_n to θ^*:

$$d(\theta^*, \theta_{n+1}) \leq \frac{q}{1-q} d(\theta_{n+1}, \theta_n). \tag{3.84}$$

Hence, we can use this method to approximate θ with arbitrary precision by applying the function $R(\theta)$ a certain amount of times to any initial value in the interval $[0, \hat{\theta}]$. The problem, however, is that we do not know how to evaluate q precisely. We conjecture that $R(\theta)$ is convex on $[0, \hat{\theta}]$, which would imply that the maximum value of $R'(\theta)$ is achieved at $\hat{\theta}$. Then, it would enable us to show that

$$q = R'(\hat{\theta}) = 1 + \frac{1-\gamma}{\gamma} \ln(1-\gamma). \tag{3.85}$$

However, we leave this question open for further investigation.

Lemma 8.

$$R(\theta) = \theta - \frac{p_v}{p_d} + o(1). \tag{3.86}$$

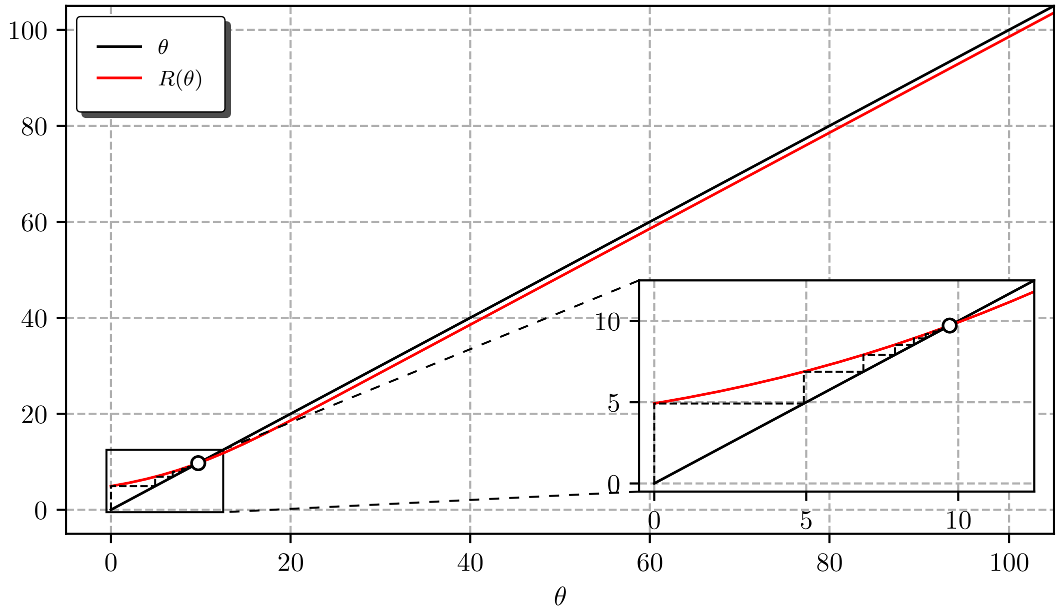


Figure 3.1: A visualization of the fixed-point iteration method applied to the model $H(H_0, p_v = 0.3, p_e = 0.49, p_d = 0.21, Y_t = 3)$, starting from $\theta_0 = 0$. We can see that $R(\theta)$ monotonously increases, it appears to be convex and asymptotically behave as a linear function; we prove the latter in Lemma 8.

Proof. According to [20], for $|\arg(1-z)| < \pi$ and positive large λ, we have

$$\begin{aligned} F(a, b; c-\lambda; z) = 1 - \frac{abz}{\lambda} + O\big(\lambda^{-2}\big) - \frac{\pi\lambda^{a+b-1}\big(z/(1-z)\big)^{\lambda-c} z}{\Gamma(a)\Gamma(b)\sin\big(\pi(\lambda-c)(1-z)\big)} \times \\ \times \left(1 + \frac{(a+b-1)(a+b+2-2c) - 2ab + 2(1-a)(1-b)z}{2\lambda} + O\big(\lambda^{-2}\big)\right). \end{aligned} \tag{3.87}$$

We observe that from this approximation, we obtain

$$\lim_{\lambda\to+\infty} \frac{\lambda}{F(a, b; c-\lambda; z)} = 0 \tag{3.88}$$

whenever $z/(1-z) > 1$.

Remind from Equation 3.41 that

$$\gamma = \frac{p_v(\mu-1) + p_e\mu}{p_v(\mu-1) + p_e\mu + p_d}. \tag{3.89}$$

Since $p_v > p_d$ and $\mu \geq 1$, we have $0 < \gamma \leq 1$ and $\gamma/(1-\gamma) > 1$. Then, we plug

$$\begin{aligned} a &= 1, & b &= 2, \\ c &= 2 + \frac{\mu(p_v + p_e)}{p_v(\mu - 1) + p_e\mu + p_d}, & \lambda &= \frac{p_d\theta}{p_v(\mu - 1) + p_e\mu + p_d}, \end{aligned}$$
$$z = \gamma$$

into Equation 3.88 to show that the right term of Equation 3.67 approaches 0 as $\theta \to \infty$, which in turn gives us Equation 3.86. $\square$

4 Results and Discussion

4.1 Simulations

In order to verify the proposed model, we ran numerous simulations with different configurations of the model to generate random hypergraphs. For example, Figure 4.1 demonstrates the empirical degree distribution of a hypergraph, simulated according to the model.[1] It also compares it with the theoretical distribution, which was calculated from the parameters of the model according to Theorems 3 and 7.

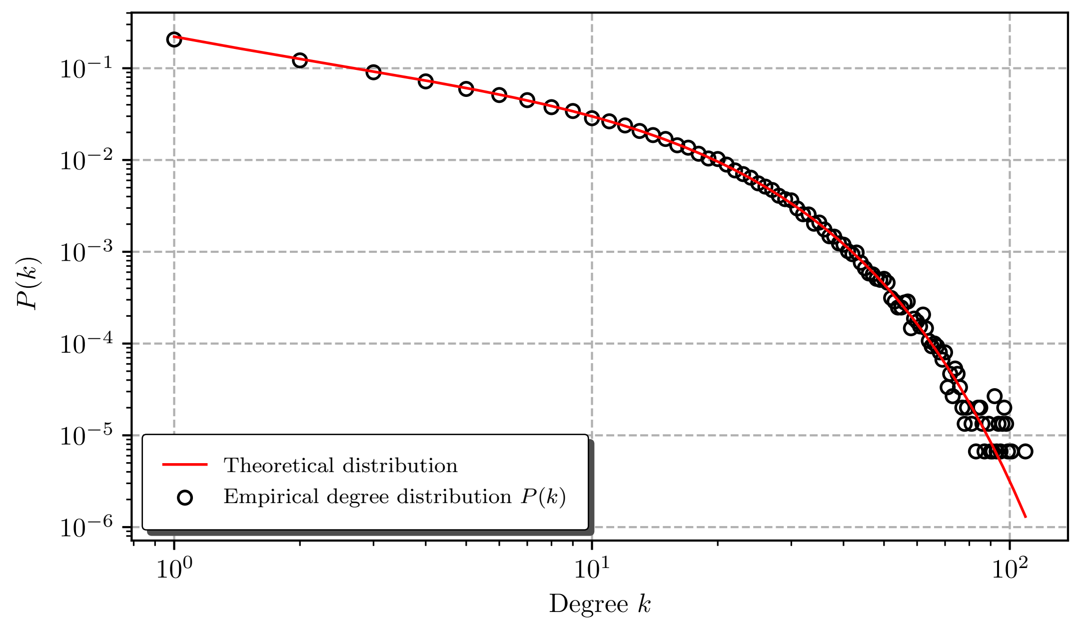


Figure 4.1: The degree distribution of $H(H_0, p_v = 0.3, p_e = 0.49, p_d = 0.21, Y_t = 3)$ after $t = 500\,000$ steps and its corresponding theoretical distribution.

In Figure 4.1, we can see that the empirical degree distribution closely corresponds to the theoretical one, which makes us believe that the assumptions we have made to prove the theorem are true.

[1]The model parameters were chosen arbitrarily. We also observed similar results with other parameters.

Figure 4.2 demonstrates the evolution of the average degree of a vertex selected for deactivation and compares it with the value of θ, which was calculated according to Theorem 7 using the fixed-point iteration method.

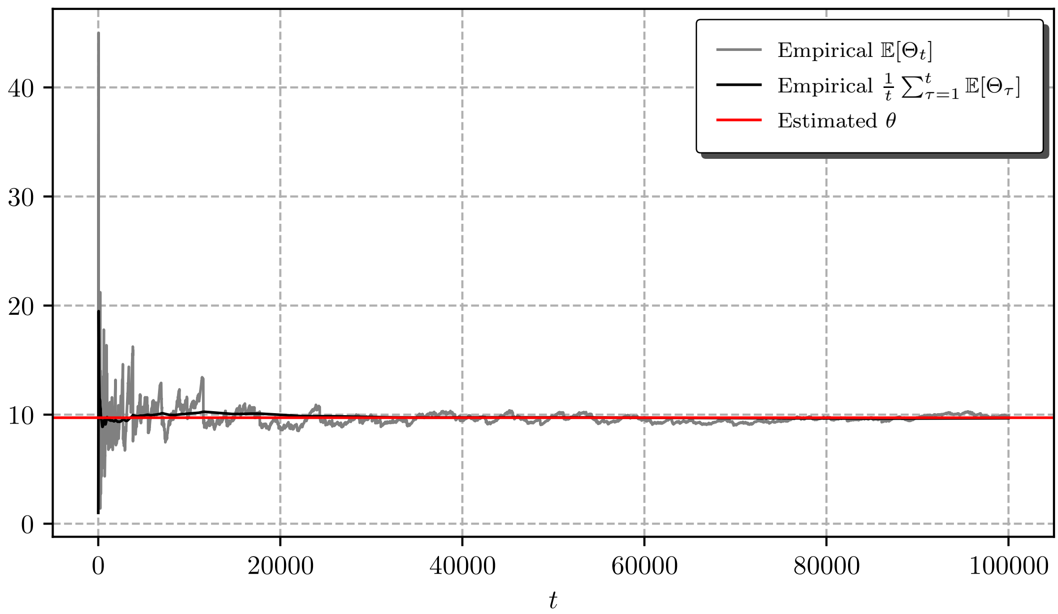


Figure 4.2: The evolution of the empirical average deactivated degree (grey and black curves) as compared to the estimated θ (red line) in a simulation of the model $H(H_0, p_v = 0.3, p_e = 0.49, p_d = 0.21, Y_t = 3)$.

From Figure 4.2, it appears that the black curve, which represents the empirical average degree of a deactivated vertex, approaches the red line, which represents the estimated value of θ. This observation supports Assumption 2 about the convergence of the average deactivated degree to some constant real number θ, as well as our method for evaluating θ using the fixed-point iteration method.

Finally, we conducted a series of simulations to generate 10 000 hypergraphs according to the model $H(H_0, p_v = 0.3, p_e = 0.49, p_d = 0.21, Y_t = 3)$ after $t = 100\,000$ steps. We used these hypergraphs to approximate the actual $\mathbb{E}[D_t]$ by the empirical mean of D_t. Figure 4.3 shows this empirical $\mathbb{E}[D_t]$ as compared to the trajectory of D_t of one of the generated hypergraphs. We can see that the empirical $\mathbb{E}[D_t]$ appears to be linear with the slope $\alpha = 0.32900254$. We then calculate the slope of the actual $\mathbb{E}[D_t]$ by using the fixed-point iteration method to compute θ and then plugging it into Equation 3.9, which yields $\alpha = 0.32904168$. Hence, we observe that the empirical estimations closely correspond to the theoretical results, and they thus support the assumptions we have previously made.

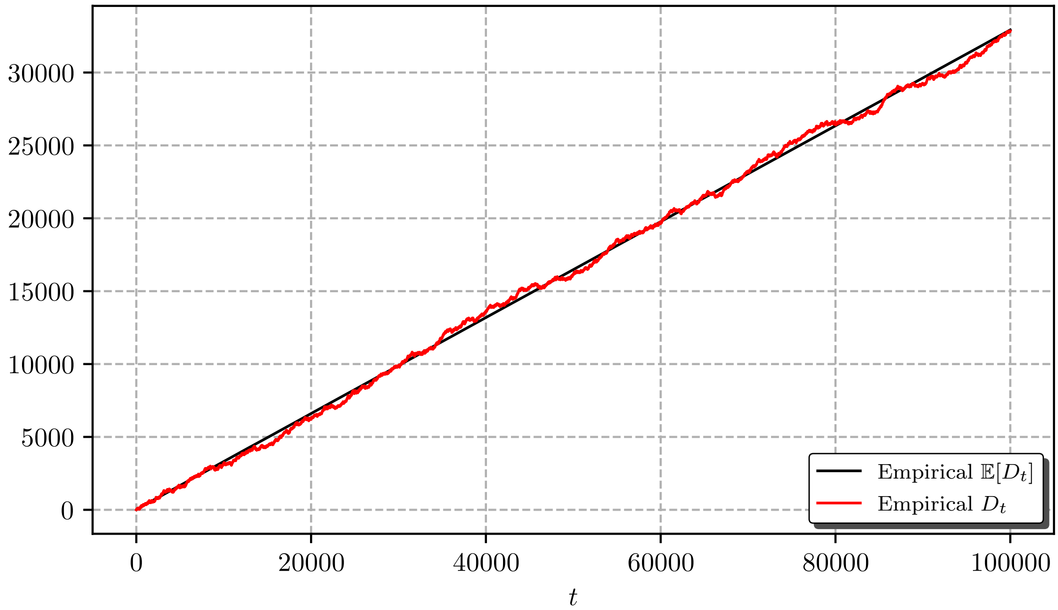


Figure 4.3: The empirical $\mathbb{E}[D_t]$ and the trajectory of D_t of a generated hypergraph.

Figure 4.4 demonstrates the concentration of D_t around its expectation. It appears that D_t can be approximated by $\mathbb{E}[D_t] + o(t)$, which thus supports Assumption 3.

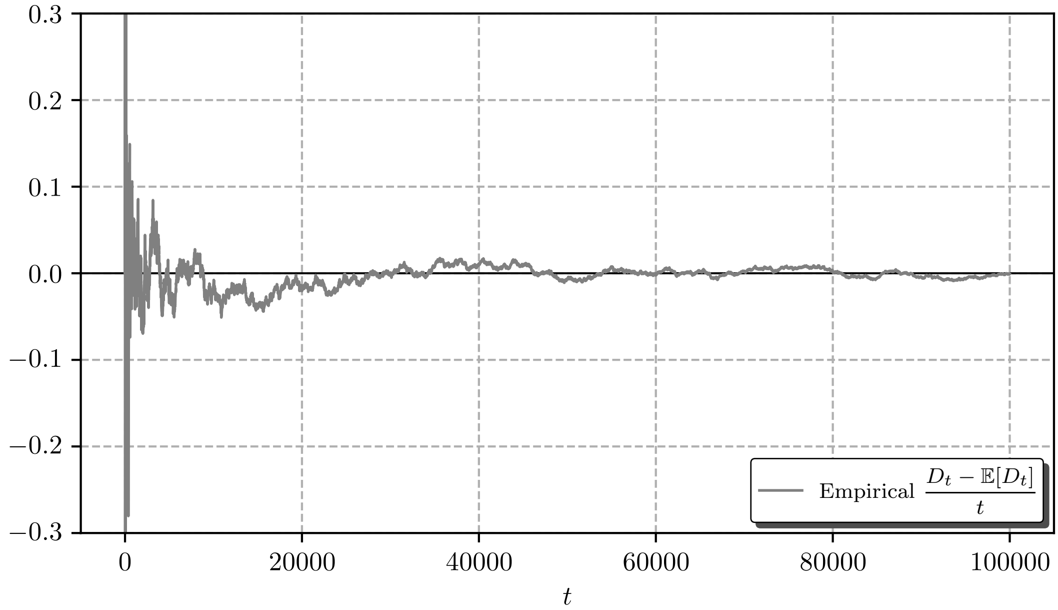


Figure 4.4: The concentration of D_t around its expectation.

4.2 Limitations

Even though we have achieved our main goal and developed a mathematical model, which generates random hypergraphs whose degree distribution follows a power-law distribution with an exponential cutoff, our model is still limited when it comes to modelling real-world collaboration networks. For example, in the network, which we discussed in Section 2.2, there were 239 414 publications and 258 145 authors. Unfortunately, we cannot apply the proposed model to such networks, since we always have that $|V_t| \leq |E_t|$. To overcome this problem, we could try the following.

- Introduce events with probabilities p_v, p_{ve}, p_e and p_d. In this case, p_{ve}, p_e and p_d would correspond to p_v, p_e and p_d in the proposed model, but we would also consider a new event, which occurs with probability p_v and adds only a vertex without a hyperedge. However, the current definition of preferential attachment will not allow vertices with degree 0 to be selected. Hence, for example, by redefining the preferential attachment function as

$$p_A(v) = \frac{\deg v + \epsilon}{\sum_{u \in A}(\deg u + \epsilon)}, \tag{4.1}$$

  where ϵ is some small positive constant, we can ensure that even vertices with degree 0 can be selected with small probability [18].

- Introduce a sequence of random variables $X = (X_1, X_2, \ldots)$, which would describe the number of new vertices to add to the hypergraph in each step. In this case, we would need to require that $X_t \leq Y_t$, for every $t \geq 1$. Then, whenever we add vertices and a hyperedge in step $t \geq 1$, we put X_t vertices to that new hyperedge of size Y_t and also preferentially select $Y_t - X_t$ active vertices from H_{t-1}.

Of course, these modifications would increase the complexity of the model and the analysis, but it could help us make this model general enough to model arbitrary collaboration networks.

5 Conclusion

In this thesis, we have presented a novel mathematical model that generates random hypergraphs whose degree distribution follows a power-law distribution with an exponential cutoff. Also, under a few assumptions, we have proved a theorem about the degree distribution of hypergraphs generated according to the proposed model, and described precise methods to evaluate the parameters of the corresponding theoretical distribution.

5.1 Further Work

There are still several open questions to address in the future work.

1. In order to prove Theorem 3, we have stated Assumptions 1–4 about some distributions associated with the model. Even though the empirical observations confirmed the validity of these assumptions, it is necessary to either prove them, or at least loosen them to make the model more rigorous.
2. In Section 3.3.3, we presented a method to calculate the value of θ by applying the fixed-point iteration method to the function $R(\theta)$. However, we could not estimate the speed of convergence of this method to the fixed point, as we did not calculate the q-Lipschitz constant of $R(\theta)$. Hence, it is important to find a method to calculate q (for example, by proving that $R(\theta)$ is convex).
3. Address the limitations, which were described in Section 4.2, and find a method to determine parameters of the model from a real-world network.

Bibliography

[1] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover, 1972.

[2] Chen Avin, Zvi Lotker, Yinon Nahum, and David Peleg. Random preferential attachment hypergraph. In *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019*, pages 398–405. ACM, August 2019. doi:10.1145/3341161.3342867.

[3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999. doi:10.1126/science.286.5439.509.

[4] Béla Bollobás and Paul Erdős. Cliques in random graphs. *Mathematical Proceedings of the Cambridge Philosophical Society*, 80(3):419–427, 1976. doi:10.1017/S0305004100053056.

[5] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web. *Computer Networks*, 33(1):309–320, 2000. doi:10.1016/S1389-1286(00)00083-9.

[6] Anna D. Broido and Aaron Clauset. Scale-free networks are rare. *Nature Communications*, 10(1), March 2019. doi:10.1038/s41467-019-08746-5.

[7] Fan Chung and Linyuan Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002. doi:10.1073/pnas.252631999.

[8] Fan Chung and Linyuan Lu. *Complex graphs and networks.* American Mathematical Society, 2006.

[9] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, November 2009. doi:10.1137/070710111.

[10] Colin Cooper, Alan M. Frieze, and Juan Vera. Random deletion in a scale-free random graph process. *Internet Math.*, 1(4):463–483, 2003. doi:10.1080/15427951.2004.10129095.

[11] Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965. doi:10.1126/science.149.3683.510.

[12] Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae Debrecen 6*, pages 290–297, 1959.

[13] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960.

[14] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '99, pages 251–262, New York, NY, USA, 1999. Association for Computing Machinery. doi:10.1145/316188.316229.

[15] Trevor I. Fenner, Mark Levene, and George Loizou. A stochastic evolutionary model exhibiting power-law behaviour with an exponential cutoff. *Physica A*, 355(2):641–656, 2005. doi:10.1016/j.physa.2005.01.007.

[16] Trevor I. Fenner, Mark Levene, and George Loizou. A model for collaboration networks giving rise to a power-law distribution with an exponential cutoff. *Soc. Networks*, 29(1):70–80, 2007. doi:10.1016/j.socnet.2005.12.003.

[17] Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and Mark E. J. Newman. Random hypergraphs and their applications. *Physical Review E*, 79:066118, June 2009. doi:10.1103/PhysRevE.79.066118.

[18] Frédéric Giroire, Nicolas Nisse, Thibaud Trolliet, and Małgorzata Sulkowska. Preferential attachment hypergraph with high modularity. Research report, Université Côte d'Azur, March 2021. arXiv:2103.01751.

[19] Panagiotis Louridas, Diomidis Spinellis, and Vasileios Vlachos. Power laws in software. *ACM Transactions on Software Engineering and Methodology*, 18(1):1–26, October 2008. doi:10.1145/1391984.1391986.

[20] Yudell L. Luke. *The special functions and their approximations*. Academic Press, 1969.

[21] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004. doi:10.1080/15427951.2004.10129088.

[22] Mark E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, July 2001. doi:10.1103/PhysRevE.64.025102.

[23] Mark E. J. Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):016131, June 2001. doi:10.1103/PhysRevE.64.016131.

[24] Mark E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132, June 2001. doi:10.

1103/PhysRevE.64.016132.

[25] Mark E. J. Newman. Coauthorship networks and patterns of scientific collaboration. In *Proceedings of the National Academy of Sciences*, volume 101, pages 5200–5205, April 2004. doi:10.1073/pnas.0307545100.

[26] Mark E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2005. doi:10.1080/00107510500052444.

[27] Scopus: Elsevier's abstract and citation database. https://scopus.com. (Online; accessed 2021-08-01).

[28] Michael P. H. Stumpf and Mason A. Porter. Critical truths about power laws. *Science*, 335(6069):665–666, February 2012. doi:10.1126/science.1216142.

[29] Jian-Wei Wang, Li-Li Rong, Qiu-Hong Deng, and Ji-Yong Zhang. Evolving hypernetwork model. *The European Physical Journal B*, 77:493–498, October 2010. doi:10.1140/epjb/e2010-00297-8.