INHOMOGENEOUS BRANCHING PROCESSES: A TALE OF TWO NETWORKS

Jimmy Jin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2017

Approved by:

Shankar Bhamidi

Andrew Nobel

Vladas Pipiras

Edward Carlstein

Jan Hannig

ABSTRACT

JIMMY JIN: Inhomogeneous Branching Processes: A Tale of Two Networks
(Under the direction of Shankar Bhamidi and Andrew Nobel)

A basic theme in probability is the use of simple approximations to study complex systems. In this thesis we leverage the humble branching process to tackle two problems on random graphs. First, we study a variant of linear preferential attachment graphs which includes a change point in the parameter set driving the attachment dynamics. Using a continuous-time branching process embedding, we show how to estimate the change point and prove its consistency via a functional central limit theorem for the number of leaves. Additionally, we analyze the long-range dependence in the evolution of the graph, showing in particular that the exponent of the degree distribution does not feel the effect of any change. Second, motivated by recent studies showing that the spread of viral content on the internet takes surprising shapes, we introduce a simple discrete-time model for social media cascades whereby the transmission probability of the cascade decays with the distance from source. We argue that such a cascade can be reasonably approximated by a generation-dependent Galton-Watson process with infinite mean, and, as a first step to understanding its growth behavior, derive a simple criteria for its extinction.

To my grandparents.

ACKNOWLEDGEMENTS

This is the most difficult portion of the thesis to write. I truly believe that everyone I have ever met played some role—however imperceptible—in setting me along the path which led to this thesis. I thank you all. But still, there are some who are more deserving of thanks than others.

Obviously, my family is at the top. A paragraph cannot possibly sum up their contribution. I can say this much though: since a very young age, I've been very curious. I was never the smartest, the most knowledgeable, or even the most hard-working. What's carried me this far is simply my ever-present desire to peek around the next corner. For fostering and preserving that spirit within me, I thank my mom, dad, and sister. To this day, the same curiosity still guides me ever forward.

In the same vein, I thank all of my teachers over the years. Special thanks are due to two of my wonderful math professors at Swarthmore College: To Phil Everson, who showed me that statistics is really just fancy guessing and that fancy guessing is actually quite fun; and to Deb Bergstrand, who more than anyone else helped me glimpse the beauty in the most abstract corners of mathematics. I'll always remember that first sense of wonder I felt as you were explaining why permutations cannot be both even and odd.

Next I would like to thank my advisors Andrew and Shankar. I won't dwell on how they trained me technically—that's a baseline requirement for any worthwhile advisor. For me, Andrew and Shankar's contribution was far deeper. Let me explain.

Somewhere in the middle of graduate school I became a little disillusioned with science. It was the combination of many factors—certain arrogant peers who seemed more occupied with self-promotion than anything else, the broken journal system, and the ever-present feeling that one's contributions are inconsequential. Some days I found it very difficult to

wake up in the morning and feel excited about contributing to such a system. I began to wonder whether the true spirit of scientific discovery was still alive out there, or whether we were all eventually just going to end up going through the motions, chasing the next hot topic which will garner us the most citations. I started to worry that my passion for research would someday wane and that I would end up just another cog in a giant paper mill.

You see, this ennui is a disease which strikes at the heart of scientific progress. Once a researcher, no matter their raw talent, loses the capacity to *dream*—to see their work as something more than just a job—then they end up as part of the problem. They spend the rest of their lives chasing prestige, citations, or the next cozy job title rather than the thing that got them in the game in the first place: the pleasure and excitement of discovering.

Therefore, I believe the true measure of a successful advisor is not their ability to train the student in technical matters but rather their ability to *inspire* in the student a deep appreciation and enjoyment in the work that they do. The benefits of good technical advice is short-lived. The benefits of a healthy passion for research last an entire lifetime.

For me, Shankar and Andrew succeeded in this wonderfully. It's difficult to put a finger on exactly what they did. But whether it was their willingness to work with me on topics far afield of their expertise, or the example they set from their own passion for tackling problems, the end result is that I **always** felt inspired and excited about learning and exploring. And that, in turn, is what kept me going even when I felt like the system was so broken that there was no point in keeping on. I can't thank them enough for that.

Thanks are also due to many individuals in my life who don't fit neatly into a category:

- To my classmates Kelly, John, Qunqun, and Ruoyu, whose positivity and humor made the Hanes basement as pleasant a place to be as it could possibly be (please don't forget about me when you're wealthy and/or successful).

- To Frances, for gratefully putting up with all the times I squatted in her house during my final semester working remotely and for generally being awesome.

- To the power trio of Alison, Sam, and Christine, who made it possible for the rest of us to keep our heads in the clouds by taking care all the stuff that keeps us connected to the ground.
- To Iain Carmichael, for hooking me up with the CourtListener data.
- To Kyle Skolfield for helping to read my thesis and ensure that there isn't too much stupidity contained within.
- To Shankar's dog Annabelle, for being really cute and not trying to eat me.
- To my cat Peachy, who surely can't read but definitely is the best stress-relieving device to ever appear on this planet.
- To Alex Valencia, for also putting with me for that one time I squatted in your house. Also Yue, and also for introducing me to...
- ...and lastly but not least, to Diana, who more than anyone else remind s me that there is much joy to be had outside of math and statistics. And that I need to eat more vegetables. I would not have made it here without you.

Finally, I would like to thank my committee—Dr. Pipiras in particular since he once remarked that the only thing he checks for in students' theses is whether or not he is mentioned in the acknowledgements—for taking the time out of their busy schedule to read and advise me on this thesis. I hope that this work makes you and the UNC STOR department proud.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

$p_{\boldsymbol{\theta}}$	limiting degree distribution of the preferential attachment graph driven by the parameter set $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$.
$D_{\boldsymbol{\theta}}$	random variable with the distribution $p_{\boldsymbol{\theta}}$, defined in 3.7.
$D_{\boldsymbol{\theta}}^{\mathsf{out}}$	$D_{\boldsymbol{\theta}} - 1$, representing a random variable with the limiting out-degree distribution.
BC, AC	before-changepoint, after-changepoint
$\mathcal{P}_\alpha(\mathcal{P}_\beta)$	the Yule-process variant driven by parameter α (β) viewed as a point process, defined in 3.4.
$N_\alpha(t)(N_\beta(t))$	number of points in $\mathcal{P}_\alpha$ ($\mathcal{P}_\beta$) which fall in the interval $[0, t]$.
$\mathcal{T}_{\boldsymbol{\theta},n}(\mathcal{T}_n)$	preferential attachment (random) tree on n vertices driven by parameter set $\boldsymbol{\theta}$.
${}_th^{(n)}$	average proportion of leaves in the graph on n vertices over all times up to t, see 3.21.
$h_t^{(n)}$	average proportion of leaves in the graph on n vertices over all times after t, see 3.22.
$D_n(t)$	scaled absolute difference between ${}_th^{(n)}$ and $h_t^{(n)}$, see 4.2.
$D(t)$	theoretical limit of $D_n(t)$ as $n \to \infty$, see 3.106.
$p_t^{(\infty)}$	the limiting proportion of leaves in the graph as $t \to \infty$ as a function of t.
$N_n(k)$	number of vertices with degree k in the random tree $\mathcal{T}_n$.
$N_n(k, m)$	number of vertices with degree k in the random tree $\mathcal{T}_n$ at the time of appearance of the mth vertex.
$\hat{N}_n(k, t)$	number of vertices with degree k in the random tree $\mathcal{T}_n$ at time nt.
$\hat{p}^n(k, t)$	proportion of vertices with degree k in the random tree $\mathcal{T}_n$ at time nt.
$\hat{p}_t^n$	$:= \hat{p}^n(1, t)$, or the proportion of leaves in the random tree $\mathcal{T}_n$ at time nt.
$\mathsf{BP}_\alpha(t)$	continuous-time branching process embedding of the preferential attachment graph driven by parameter α, see 3.4.1.
$\lvert\mathsf{BP}_{\boldsymbol{\theta}}(t)\rvert$	number of individuals in $\mathsf{BP}_{\boldsymbol{\theta}}(t)$ at time t.
$\mathsf{BP}_{\boldsymbol{\theta}}^n(t)$	continuous-time branching process embedding of the preferential attachment graph driven by parameter set $\boldsymbol{\theta}$, allowed to grow until size n.

Υ_n	amount of time after the change point until the process reaches size n in the continuous-time embedding of the preferential attachment with change point model.
a	limiting expectation of Υ_n, equal to $1/(2+\beta)\log(1/\gamma)$.
Age	the age (in the continuous-time embedding of preferential attachment) of a vertex born after the change point by the time the process reaches its final size.
$k_n(s)$	cumulant generating function for the offspring of an individual in generation n.
$k^{(n)}(s)$	nth functional iterate of the cumulant generation functions $\{k_m\}_{0\leqslant m<n}$.
$h_n(s)$	functional inverse of $k_n(s)$.
$h^{(n)}(s)$	functional inverse of $k^{(n)}(s)$.

CHAPTER 1

Introduction

In the theory of random graphs, most of the answers can be guessed using the heuristic that the growth of the cluster is like that of a branching process.

-Rick Durrett, *Random Graph Dynamics*

Graphs and networks are more relevant than ever. Whether in statistics, computer science, physics, or economics, the random graph has seen its role come front and center in the past decade as the world has grown more interconnected. For the statistician or probabilist, this is both the best of times and the worst of times. Best—because of an explosion of real-world data with which to drive new ideas and models. But also worst—because real-world phenomena does not always fit into neat, simplified mathematical models. Increasingly, we are preoccupied with inventing better, more accurate models to fit what we observe in reality.

However, this thesis is not an effort to do that. The reader will find no pretense in this thesis of claiming to outperform an existing method for modelling a network or for estimating some parameter of a stochastic block model. Rather, it is a testament to the power of a simple tool, the humble branching process, to achieve deep insights. In this thesis, we show that this simple stochastic process, familiar to any sufficiently advanced undergraduate, can reveal deep insights when properly employed.

Early on in the author's graduate career his advisor assigned some reading from [46]. In Chapter 2 there is the following theorem:

Theorem (2.4.1). *Suppose we have an Erdős-Rényi random graph with $\lambda > 1$. If we pick two points at random from the giant component, then*

$$\frac{d(x,y)}{\log n} \to \frac{1}{\log \lambda} \quad \textit{in probability.}$$

The explanation for the theorem was the following:

> *The answer in Theorem 2.4.1 is intuitive. The branching process approximation grows at rate λ^t, so the average distance is given by solving $\lambda^t = n$, i.e., $t = (\log n)/\log \lambda$.*

To someone completely new to random graph theory working through a book clearly written for other probabilists, seeing a familiar structure like *branching processes* is like happening across a road after being lost in the woods for several days. Branching processes became a path which the author could follow to delve deeper into random graph theory without fear of losing his way.

And to the author's surprise, they have never stopped serving that purpose. Indeed, if one squints hard enough, branching processes can be found in many random graph models. In this thesis, we explore two important topics in random graph theory through the lens of branching processes.

1.1 1.1. Summary of thesis

1.1 1.1.1. Preferential attachment with change point

> **Question:** Suppose we have a network growing over time, controlled indirectly by a parameter governing how new nodes join the network. If that parameter experiences a sudden change, how can we estimate it?

Preferential attachment is one of the most important models not only in random graph theory, but also in sociology and economics. It's one of very few *temporal* network models

which is simple, yet capable of producing many characteristics seen in real-world graphs—namely, a power-law degree distribution. We propose a simple change point variant of this model, investigate some non-trivial ramifications, and then show how one can estimate the change point.

The basic preferential attachment graph is grown by the following scheme:

Start with a single vertex ρ at time $m = 1$ (this vertex will be referred to as the *root* or the original progenitor of the process). Fix a parameter $\alpha > -1$. At each discrete-time point $1 < m \leqslant n$ a new vertex enters the system with a *single edge*[1] which it will then connect to a pre-existing vertex. The vertex connects to a pre-existing vertex v with probability proportional to the current degree of v $+\alpha$.

Now consider the same model but with a change point in the attachment parameter α. Fix two attachment parameters $\alpha, \beta > -1$, a change point parameter $\gamma \in (0, 1)$, and a system size $n > 1$. The model does preferential attachment as before, but now the attachment dynamics changes after time $\lfloor n\gamma \rfloor$ namely

(a) For time $0 < m \leqslant \lfloor n\gamma \rfloor$, the new vertex entering the system at time m connects to pre-existing vertices with probability proportional to their current out-degree $+1 + \alpha$.

(b) For time $\lfloor n\gamma \rfloor < t \leqslant n$, the new vertex connects to pre-existing vertices with probability proportional to their current out-degree $+1 + \beta$.

We immediately have two questions:

1. How does the change point affect the aggregate characteristics of the graph?

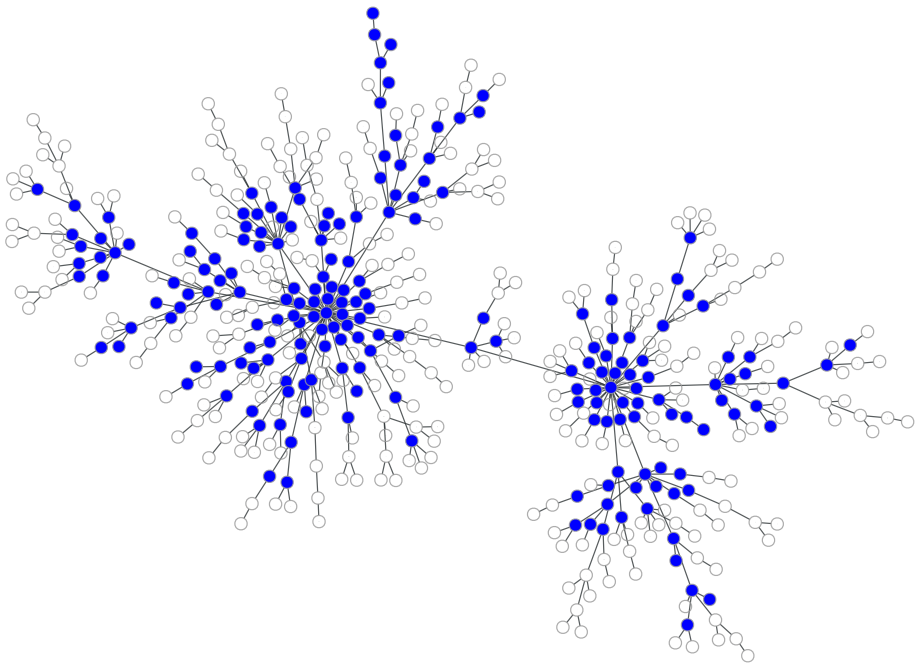
2. How can we estimate the change point γ?

Our answer to question (1) is summarized in Theorem 3.2.1, which says that, at least with respect to the power-law exponent of the degree distribution, the answer is actually "not that much."

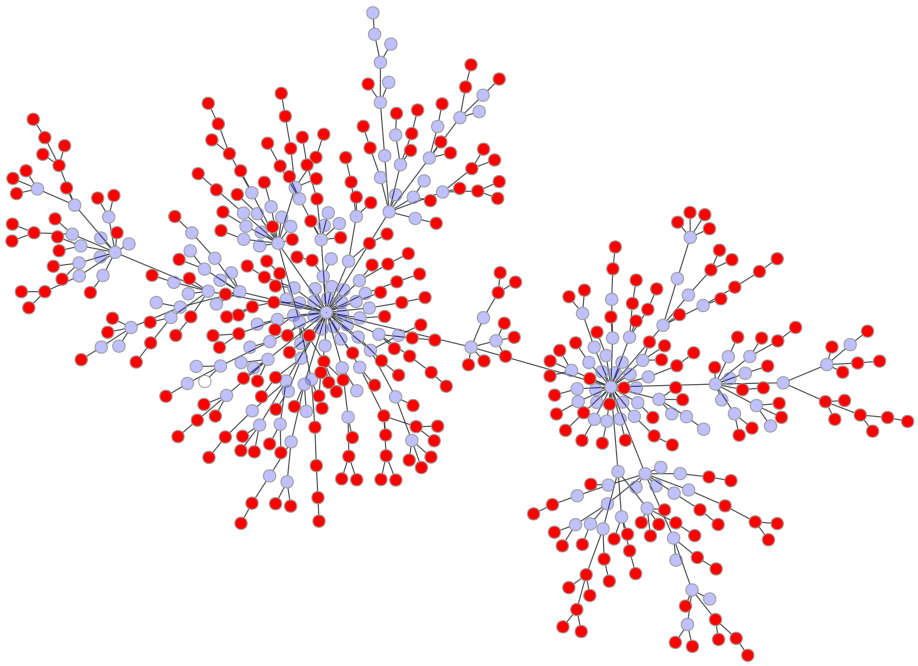
[1]Throughout this chapter we will consider the simplest case where the network at each stage is a tree. The methodology can be generalized.

Figure 1.1: The evolution of a preferential attachment graph with change point at $\gamma = 0.4$.

(a) Evolution from $t = 0.0$ until $t = 0.4$ with $\alpha = -0.5$.



(b) Evolution from $t = 0.4$ until $t = 1.0$ with $\beta = 30$.



Theorem 3.2.1. *Fix some parameter set for the change point model* $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$ *and* $k \geqslant 1$. *There exists an integer-valued random variable* $D_{\boldsymbol{\theta}}$ *such that as* $n \to \infty$, *the proportion of vertices in the graph with degree* $k := N_n(k)/n$ *satisfies*

$$\frac{N_n(k)}{n} \xrightarrow{\text{P}} \mathbb{P}(D_{\boldsymbol{\theta}} = k), \quad \text{as } n \to \infty$$

However, *there exist constants* $0 < c < c'$ *such that for all* $k \geqslant 1$

$$\frac{c}{k^{\alpha+2}} \leqslant \mathbb{P}(D_{\boldsymbol{\theta}} \geqslant k) \leqslant \frac{c'}{k^{\alpha+2}}. \tag{1.1}$$

Notably, the scaling in equation 1.1 does not involve either β or γ. So no matter how big the jump between α and β is, and no matter how early the change point γ occurs, the limiting power-law scaling still only depends on the initial parameter value α. In Chapter 3 we will develop the correct continuous-time branching process framework which will lay bare why this is.

The second thrust of our analysis focuses on trying to answer question (2). How can we best estimate the change point γ?

Our proposal is based on counting leaf nodes. We think this approach is not just elegant, but also potentially extensible to non-preferential-attachment-like networks because it does not explicitly employ the likelihood function for preferential attachment. The rationale is as follows: since the attachment parameter directly affects the chance that new vertices attach to leaves (see Figure 1.1), one ought to be able to feel the effect of the change point through this statistic.

The basic idea behind our estimator is to scan through all time points t in the history of the graph and calculate the difference between:

$$_{t}h := \text{ the average proportion of leaves in the graph over all times up to } t \tag{1.2}$$

$$h_t := \text{ the average proportion of leaves in the graph over all times after } t \tag{1.3}$$

Then, it makes sense to estimate the change point $\hat{\gamma}$ to be the argmax of this function. Actually, we'll need to scale the difference first—the precise scaling tells us a lot about the unique nature of this change point problem—but we defer a deeper discussion about this point to Subsection 3.3.2 (see Figure 3.2 for an example of $D_n(t)$).

$$D_n(t) := (1-t)|_t h^{(n)} - h_t^{(n)}|, \qquad t \in [\varepsilon, 1] \tag{1.4}$$

Here's a walkthrough of how we will prove this estimator is consistent. It turns out that we can calculate what the limit $D(t)$ of this function should be quite easily: it's constant up to time $t = \gamma$, and then decreases smoothly towards 0, which it achieves at $t = 1$. Therefore if we can just consistently estimate the point at which it starts decreasing, then we are done.

To do this we will need to understand the order of the fluctuations of $D_n(t)$ around $D(t)$. We accomplish this via a functional central limit theorem for the proportion of leaves. Let $\hat{N}_n(1,t)$ be the number of leaves in the graph of size n *at time* $t \in [0,1]$. Then:

Theorem 3.2.3. *Let $p_t^{(\infty)}$ be a function in t describing the limiting proportion (as $n \to \infty$) of leaves in the with-change point graph. Consider the process of re-centered and normalized number of leaves*

$$G_n(t) := \frac{\hat{N}_n(1,t) - ntp_t^{(\infty)}}{\sqrt{n}}, \qquad 0 \leqslant t \leqslant 1, \tag{1.5}$$

with linear interpolation between time points. Then as $n \to \infty$, $G_n \xrightarrow{\text{w}} G$ where G is a tight Gaussian process on $[0,1]$ and $\xrightarrow{\text{w}}$ denotes weak convergence on $D[0,1]$ equipped with the usual Skorohod metric.

It follows therefore:

Lemma 3.4.28. *Fix $\varepsilon > 0$. Then*

$$\sup_{t \in [\varepsilon, 1]} |D_n(t) - D(t)| = O_P\left(\frac{1}{\sqrt{n}}\right)$$

Now we can define our estimator quite naturally by:

$$\hat{\gamma}_n := \max\{t : t \in \mathcal{M}_n\}. \tag{1.6}$$

Where the set $\mathcal{M}_n$ is the collection of points t for which the corresponding function value $D_n(t)$ is within $\log n/\sqrt{n}$ of the maximum of the function:

$$\mathcal{M}_n := \left\{ t \in [\varepsilon, 1] : |D_n(t) - \max_{t \in [\varepsilon,1]} D_n(t)| \leqslant \frac{\log n}{\sqrt{n}} \right\} \tag{1.7}$$

Then it follows finally that

Theorem 3.2.4. *Assume that the change point $\gamma > \varepsilon$. Then $\hat{\gamma}$ is consistent and, in fact,*

$$|\hat{\gamma}_n - \gamma| = O_P\left(\frac{\log n}{\sqrt{n}}\right) \tag{1.8}$$

In Chapter 4 we use simulations to examine the performance of this estimator and also to build some more intuition about how it behaves and how it might be extended to other settings. In addition we take the philosophy of looking at functions of the graph history and run with it on two real-world temporal networks to see what lessons we might take from it for future work.

1.1 1.1.2. Decreasing cascades and thinned branching processes

> **Question:** Is a simple, discrete-time branching process model enough to capture the real-world behavior of information cascades?

The next object of study in this thesis is the *cascade*. A *cascade* can mean a lot of different things depending on the context. Loosely, a cascade is a process occurring on a graph which spreads across vertices in a way such that affected vertices trigger neighboring vertices in some way in either continuous or discrete time.

The real-world phenomenon inspiring our study is the *retweet cascade* endemic to Twitter. Twitter is a social media platform whereby users can send out (or "tweet") short

messages to other users who follow them (the user's "followers"). When a follower receives a tweet, they can either read it passively or choose to pass it on to *their* followers ("retweeting"). And clearly, if an original tweet is interesting enough, then it can trigger a large cascade flowing across the entire Twitter social network.

Generally, we refer to these large cascades as *viral*. And lately, empirical studies have revealed an interesting of viral cascades: they do not all look the same. Empirical studies have shown that viral cascades can have different shapes, from a very wide but shallow tree sometimes called a *broadcast*, to a very narrow but deep chain of retweets.

As it turns out, conventional branching process theory does not allow for some of the possibilities. Our goal in this chapter is to take the first step towards achieving these differently-shaped viral cascades using as parsimonious a branching process model as possible.

The cascade we want to model in this chapter is a simple discrete-time models which we call the *decreasing cascade* model, chosen because of how closely it mimics sharing dynamics on modern social networks:

Definition 1.1.1. ***The decreasing cascade***

Let G be a graph with vertex set V. A decreasing cascade *explores G in discrete time through a set of* active *nodes, tracing a tree structure in the following way.*

Let $\{p_n\}_{n\geqslant 1}$ be a decreasing sequence of probabilities.

1. *At time 1, start with one infected node.*

2. *At time n, active nodes infect their unexplored neighbors independently with probability p_n. In other words, vertices adjacent to currently-infected nodes become infected with probability p_n.*

3. *Once a node is finished infecting its neighbors or has failed to become infected, it cannot infect any more nodes.*

4. *The set of nodes which were successfully infected during time n then becomes the set of active nodes of time $n+1$.*

5. Repeat until the cascade dies out or reaches the entire graph.

As it turns out, if we care only about tracking the *number* of infected nodes in each discrete-time step of such a cascade, then we can re-imagine the decreasing cascade as a sort of two-step branching process. An individual in the process (infected vertex) has a certain number of offspring (the number of neighbors of the vertex) which are then "thinned" binomially to produce their final contribution to the next generation (the number of *infected* neighbors of the vertex).

Definition 1.1.2. ***The thinned branching process***

Let $\{p_n\}_{n\geqslant 1}$ be a decreasing sequence of probabilities and let us start with 1 individual in the system at time 1.

At time n, each individual X gives birth independently to a random number of offspring W, which is distributed according to some offspring distribution F (common to all time steps).

Then, perform Binomial thinning on the offspring by letting only $Y \sim Binomial(W, p_n)$ survive until the next generation. Repeat indefinitely or until extinction of the process.

Since many real-world networks have power law degree distributions, we will constrict ourselves to the case where the decreasing cascade is flowing on a scale-free network with degree distribution tail exponent $\in (1, 2)$[2]. It turns out that because of size-biasing, this is equivalent to setting the offspring distribution F to a power-law distribution with tail exponent $\in (0, 1)$, implying that the number of infected individuals in each generation is a random variable with infinite mean.

This takes us into uncharted territory. There has been plenty of research into infinite mean Galton-Watson processes and plenty of research into inhomogeneous branching processes with finite mean.

We contribute one basic result to the literature, which is:

[2]We use the term *tail exponent* to emphasize that it is the exponent α as in $\mathbb{P}(X > k) \propto k^{-\alpha}$

Theorem 5.3.3. *Suppose that F is a probability distribution satisfying*

$$1 - F(x) \sim \frac{C}{x^{\alpha}}, \quad \alpha \in (0, 1) \tag{1.9}$$

Then a branching process with binomial thinning of such an offspring distribution extinguishes with probability 1 if and only if the thinning probabilities $\{p_n\}_{n \geqslant 1}$ satisfy

$$-\sum_{k=1}^{n} (1/\alpha)^{-k} \log p_k \to \infty \quad \text{as } n \to \infty$$

And we close by discussing some next steps which should allow us to prove a Kesten-Stigum-type result for these branching processes.

CHAPTER 2

Background and literature

2.1 2.1. The growth of branching populations

> *Let us imagine objects that can generate additional objects of the same kind; they may be men or cats reproducing by familiar biological methods, or neutrons in a chain reaction. An initial set of objects, which we call the 0-th generation, have children that are called the first generation; their children are the second generation, and so on. The process is affected by chance events.*
>
> *-T.E. Harris,* The Theory of Branching Processes

The findings in this thesis are actually just glorified studies of very simple branching processes. Therefore, a thorough understanding of our new results requires a familiarity with some foundational results.

In particular, we will rely heavily on results describing the *growth of supercritical processes*—i.e. processes which have some probability of surviving forever. Let us take a short refresher course through the literature on this topic.

2.1 2.1.1. The Dummies' guide to the Kesten-Stigum theorem

The building block of all branching process theory is the Galton-Watson process.

Suppose that in a reproducing process like in the quote above we keep track only of sizes of the successive generations and not the times of reproduction or their family relationships. Then we can write down a formal definition of the branching process quite easily. Denote by $Z_0, Z_1, Z_2, \ldots$ the numbers of individuals in the 0-th, first, second, ... generations of the process. In what follows, we shall always assume that $Z_0 = 1$ unless otherwise stated.

Let $\mathbb{P}$ denote the probability measure of the process. The probability distribution of Z_1 is prescribed by putting $\mathbb{P}(Z_1 = k) = p_k$, for $k = 0, 1, \ldots$ with $\sum_k p_k = 1$ where p_k is interpreted as the probability that an object existing in the nth generation has k children in the $(n+1)$th generation.

If the process adheres to the following two assumptions,

1. **Homogeneity**: p_k does not depend on the generation number n
2. **Independence**: All individuals beget offspring independently of each other

then it is a *Galton-Watson branching process ("GWBP") with offspring distribution* $\mathbb{P}$. The most basic result in GWBP theory tells us that survival of the family line of Z_0 depends critically on μ, the mean of the offspring distribution $\mathbb{P}$.

Theorem 2.1.1. *If* $\mu \leqslant 1$*, then* $\mathbb{P}(\exists n : Z_n = 0) = 1$ *except for the degenerate case of* $p_1 = 1$*. Otherwise when* $\mu > 1$*,* $\mathbb{P}(\exists n : Z_n = 0) < 1$*.*

When the process is guaranteed to die out, we say it is *critical* or *subcritical*, depending on whether $\mu = 1$ or $\mu < 1$ respectively. When the process may survive indefinitely, which happens when $\mu > 1$, we say it is *supercritical*. All in all, the extinction result is not very surprising. It says that if each individual in the system has at most one child on average, then the family line will eventually die out. Most people can believe this.

Where it starts to get interesting is in the analysis of the growth rate of GWBPs. First off, branching processes do not stay stable—they either go the way of the Dodo, or they go the way of *Homo sapiens*:

Theorem 2.1.2. *For any (non-degenerate) regime,* $\mathbb{P}(\lim Z_n = 0) + \mathbb{P}(\lim Z_n = \infty) = 1$*.*

But then after that, precise long-run growth analysis depends sensitively on which regime we are in. So just how does one begin to quantify the growth rate of $\{Z_n\}_{n \geqslant 0}$, a sequence of random variables? By comparing it to a reference sequence. As it turns out, the sequence

$\{\mu^n\}_{n\geqslant 0}$ works pretty well, because *at least when* $\mu < \infty$, the sequence $\{M_n\}_{n\geqslant 0}$ defined by

$$M_n := \frac{Z_n}{\mu^n}, \quad n \geqslant 0 \tag{2.1}$$

has $\mathbb{E}\, M_n = 1$ for all n, so that, at least in expectation, μ^n matches Z_n.

But this is not the end of the story. $\{M_n\}_{n\geqslant 0}$ is a non-negative martingale adapted to the filtration generated by $\{Z_n\}_{n\geqslant 0}$, so by the martingale convergence theorem it converges almost surely to an a.s. finite limit M as $n \to \infty$. Surely, if μ^n truly measures the rate of increase of the GWBP then we will have $\mathbb{E}\, M = 1$ as well.

Surprisingly, this is not exactly the case. Indeed, if $\mu \leqslant 1$, then the GWBP is guaranteed to extinguish and $M = 0$ a.s. Therefore $\mathbb{E}\, M_n \not\to \mathbb{E}\, M$. But what about the supercritical regime?

It turns out that whether or not μ^n describes the growth rate of Z_n in the limit depends on whether the offspring distribution satisfies an $X \log X$ integrability condition. This is the content of the Kesten-Stigum theorem:

Theorem (Kesten-Stigum). *Let X stand for a random variable with the offspring distribution $\mathbb{P}$. The following are equivalent:*

1. $\mathbb{E}(M) = 1$
2. $\mathbb{E}(X \log^+ X) < \infty$
3. $\mathbb{P}(M > 0) > 0$

Additionally, we would like to draw the reader's attention to condition (3) in the above theorem. By the martingale convergence theorem we know that $M < \infty$ a.s., so that the scaling μ^n cannot fail by being too *slow*. What this seems to suggest, then, is that the scaling μ^n can sometimes be "too fast" and overwhelm Z_n, forcing $M = 0$. But actually[1]:

[1]For an elegant explanation of this, see Remark 1 in chapter 1.8 of [56].

Corollary 2.1.3.

$$\mathbb{P}(M = 0 \,|\, Z_n \to \infty) = 0$$

So on the set of non-extinction, $Z_n \sim M\mu^n$ and the different points inside the set $\{0 < M < \infty\}$ describe multiplicative "shifts" of the more-or-less parallel growth of Z_n and μ^n.

To summarize, branching processes are sensitive. They either become so large as to grow indefinitely, or they die out. And in the case with endless growth, there is a very precise rate of growth which is "correct," and even then we require special conditions on the branching process to achieve it. In the second part of this thesis, we shall try to derive similar conditions on a similar, but wilder type of branching process.

For a concise summary of the main results in the study of the limiting behavior of GWBPs, we direct the interested reader to [76].

2.1 2.1.2. Inhomogeneous and infinite-mean processes

We just showed a conventional martingale analysis of the GWBP. But there is another line of attack to branching process theory using probability generating functions.

Returning to the notation of the previous subsection, the probability generating function (or *pgf*) of a probability distribution $\mathbb{P} = \{p_k\}_{k \geqslant 0}$ supported on $\mathbb{Z}^+$, is

$$f(s) := \sum_{k=0}^{\infty} p_k s^k$$

The beauty of pgfs is that there are a couple different ways of looking at them. For one, pgfs are power series with coefficients in $[0, 1]$, so all the usual theorems apply. Secondly, if $X \sim \mathbb{P}$ then $f(s) = \mathbb{E}(s^X)$. Immediate from these observations are results like

1. $f(s)$ is continuous, and specifically continuous from the left at $s = 1$.

2. $f'(s) = \sum k s^{k-1} p_k = \mathbb{E}(X s^{X-1})$

3. $\mu = \mathbb{E}\, X = f'(1)$

But also, pgfs are especially relevant with respect to branching processes because we can compose individual offspring distribution pgfs to get pgfs of total population sizes:

Theorem 2.1.4. *Let f denote the pgf of the offspring distribution in a GWBP. Also let $f_n := \mathbb{E}(s^{Z_n})$ denote the pgf of the distribution of Z_n. Then*

$$f_{n+1}(s) = f(f_n(s))$$

This gives us an easy way to analyze distributions of future generations of the process. And this can be just as useful as the martingale analysis when examining the behavior of the process in the limit. For example, the extinction theorem 2.1.1 is usually proved using pgfs and, as a side benefit, often make explicit calculations easy:

Theorem 2.1.5. *For a supercritical GWBP, the probability of extinction is a solution to the fixed point equation $f(s) = s$.*

Proof. Write $Q := \{\exists n : Z_n = 0\}$ for the extinction event so that $\mathbb{P}(Q)$ is the probability of extinction. Also let $Q_n := \{Z_n = 0\}$.

Clearly, $Q_n \subset Q_{n+1}$ and so $\mathbb{P}(Q_n) \uparrow \mathbb{P}(Q)$ as $n \to \infty$. But also, $\mathbb{P}(Q_n) = f_n(0)$. Therefore by Theorem 2.1.4, $f_{n+1}(0) = f(f_n(0))$ which means that $\mathbb{P}(Q_{n+1}) = f(\mathbb{P}(Q_n))$. Then since f is continuous,

$$f(\mathbb{P}(Q)) = f(\lim_{n\to\infty} f_n(0)) = \lim_{n\to\infty} f(f_n(0)) = \lim_{n\to\infty} f_{n+1}(0) = \mathbb{P}(Q)$$

■

So how does the picture change if we let the offspring distributions depend on the generation number? In the literature, there are two ways to let this happen.

1. **Varying environment:** Let the offspring distributions still be deterministic, but vary by generation.

   That is, let ϕ_n denote the pgf of the offspring distribution for individuals in the $(n-1)$th generation. If $\{\phi_n\}_{n\geqslant 0}$ is deterministic, then we are in a varying environment.

2. **Random environment:** Let the offspring distributions be random across generations, but iid (stationary, ergodic).

   That is, let $\{\zeta_n\}_{n\geqslant 0}$ be a sequence of iid (stationary, ergodic) random "environmental variables" in some space Θ, where we associate with each point $\zeta \in \Theta$ a pgf ϕ_ζ. If ϕ_{ζ_n} is the pgf of the offspring distribution for individuals in the $(n-1)$th generation, then we are in a random environment.

The case of varying environment has been considered since the dawn of branching processes, and a concise summaries of main results can be found in [64] or [51]. In general, because there is no additional randomness in these processes their behavior can be well-understood so long as one can deal with the analysis of the generation functions.

First, varying environment processes behave similarly to GWBP in many ways. For example, a Kesten-Stigum theorem holds for a class of supercritical varying environment processes. Let μ_j be the mean of the offspring distribution of the jth generation and call the process *uniformly supercritical* if

$$\prod_{j=k}^{n+k-1} \mu_j \geqslant Bc^n \quad \text{for some } B > 0, c > 1, \text{ and all } n, k \geqslant 0$$

Also let X_n represent the random number of offspring of an individual in the nth generation. Then:

Theorem 3 from [45]). *If the branching process is uniformly supercritical and is dominated in the sense that there exists a random variable X with $\mathbb{E}\,X < \infty$ such that*

$$\mathbb{P}(X > x) \geqslant \mathbb{P}(X_n/\mu_n > x) \quad \textit{for all } x,$$

then there exists a sequence of constants c_n such that Z_n/c_n converges to an a.s. finite random variable W with $\{W = 0\} = \{Z_n \to 0\}$.

But there are some surprising differences in contrast to GWBPs. We have seen in the discussion of the Kesten-Stigum theorem that, on the set of non-extinction, a supercritical GWBP has essentially only one rate of growth (up to multiplicative shifts): μ^n. As shown in the theorem above, this happens to be true for a large class of branching processes in varying environment. But it is not always so. For example, the authors of [77] construct a branching process in varying environment which is supercritical and grows like 2^n on one part of the sample space and m^n with $m > 4$ on another part, both with positive measure. We shall discuss these points further later on in the thesis.

The case of random environment was first introduced in [100] where the sequence $\{\zeta_n\}_{n\geqslant 0}$ was taken to be iid. Their results were later extended to any stationary, ergodic sequence in [8] and [7] where extinction criteria and limit theorems for the process Z_n were developed.

The main takeaway from these papers is basically that under some reasonable conditions on $\{\zeta_n\}_{n\geqslant 0}$, we can see the same usual behavior of the ordinary GWBP, with slight obvious modifications. For the sake of brevity we leave the specifics to the reference.

Theorem 1 from [7]. *Under some mild assumptions about the environmental process $\{\zeta_n\}_{n\geqslant 0}$ including an $X \log X^+$ condition, essentially the same results as in the Kesten-Stigum theorem for Galton-Watson processes apply.*

We note at this point that the entire line of work above all share a common assumption: the mean of the offspring distributions is finite. The other line of the work on branching processes concerns the Galton-Watson process with infinite mean. In this literature the

offspring distributions are not taken to be varying, but have infinite mean. While technically these are just supercritical GWBPs, the conditions of the Kesten-Stigum theorem are not at all satisfied, so the limiting behavior is markedly different.

In [94], the authors adapt techniques from the study of finite-mean supercritical branching processes in [95] and [58] to characterize infinite-mean branching processes as either *regular* or *irregular* depending on their limiting growth behavior.

Recall from the Kesten-Stigum theorem that, for finite-mean processes, the probability that the martingale limit $\lim_{n\to\infty} Z_n/\mu^n = M$ is not zero is positive if and only if the $X \log^+ X$ condition is satisfied. Therefore in the infinite-mean realm, we say a process is *regular* if for any sequence of positive constants $\{c_n\}_{n\geqslant 0}$ for which $\lim Z_n/c_n$ a.s. exists, $\mathbb{P}(\lim Z_n/c_n = 0 \text{ or } \infty) = 1$.

However, just like how branching processes in varying environment surprisingly can display growth at two different rates, infinite-mean branching processes display interesting exceptions to the finite-mean behavior.

Theorem [94]. *There exist infinite-mean Galton-Watson processes such that for some positive deterministic sequence* $\{c_n\}_{n\geqslant 0}$*, the martingale limit* $M := \lim_{n\to\infty} Z_n/c_n$ *has* $\mathbb{P}(M > 0) > 0$.

Call these the *irregular* processes. In [94], it is also shown that for all regular processes there exists a slowly-varying function $U(\cdot)$ such that $U(Z_n)/e^n$ converges to a non-degenerate limit. In Chapter 5 we take the first step towards investigating these behaviors for a new, related type of branching process.

2.1 2.1.3. Continuous time branching processes

Say that in a branching process we now want to keep track of the birth, death, and reproduction times of each individual. Enter the continuous time branching process.

We saw that discrete-time branching processes are generally defined by the offspring distribution(s) of the system. If the distributions are the same for every individual, then

we have a GWBP. If they vary by generation, then we have a branching process in either a varying or random environment.

A continuous time branching process requires a bit more definition. In the most general form, we start by associating with every individual x in the system two processes:

1. λ_x (the *life-length* of x): a (possibly infinite) non-negative random variable.

2. ξ_x: (the *reproduction process* of x): a point process on $\mathcal{N}(\mathbb{R}^+)$, the space of integer or infinite-valued positive measures on $\mathbb{R}^+$ that are finite on bounded sets.[2]

It turns out that carefully defining these two processes allows us to subsume almost every other classical branching process. We will not dive deeper into the probabilistic setup here. Instead, let us build up some intuition.

One can think of ξ_x as tracing out the "timeline" of births of the individual x over the infinite time horizon, starting at the time of x's own birth[3]. This also means that in order to come into agreement with the physical reality that most things in the universe cannot reproduce after they cease to exist, we will enforce the assumption that the probability that ξ_x puts any mass after λ_x is zero, or:

$$\mathbb{P}(\xi_x((\lambda_x, \infty)) = 0) = 1 \tag{2.2}$$

Unless otherwise specified, all following statements will be conditional on no children after death. In general, note that,

1. These processes are homogeneous in the sense that we generally assume the pairs (λ_x, ξ_x) to be iid *across individuals* say with probability distribution Q, a measure on

[2]We will adhere to the usual point process notation that, for an interval A on the real line, $\xi(A)$ = the number of points inside A.

[3]In some places the ξ_x process is indexed relative to absolute time, so that time 0 represents the start of the entire branching process. In that case, if σ_x represents the birth time of x then $\xi([0, \sigma_x)) = 0$. For ease of exposition here, though, we shall let time 0 represent the time of x's birth.

the space $(\mathbb{R}^+ \times \mathcal{N}(\mathbb{R}^+))$. But they are quite inhomogeneous in the sense that the reproduction ξ_x *within* each individual's lifetime is generally not uniform.

2. Any general branching process can be easily collapsed into the classic discrete-time picture by ignoring λ_x and recording only $\xi_x([0, \lambda_x])$ instead of all of ξ_x.

Aside from this basic assumption though, there are a plethora of possible models to investigate.

Example 2.1.6. *(Bellman-Harris processes)*

A Bellman-Harris process *is a continuous-time branching process where an individual's lifespan and the number of children they bear over their lifetime are independent.*

That is, if for each individual x, λ_x is independent of ξ_x, then the branching process is a Bellman-Harris process.

Example 2.1.7. *(Splitting processes)*

A process where individuals are replaced by their offspring, in effect "splitting" into a certain number of other individuals, is called a splitting process*. In other words, individuals cannot give birth more than once.*

That is, if ξ_x gives mass to only one *random point ν, then the branching process is a splitting process.*

If further we have $\mathbb{P}(\xi_x(\{\lambda_x\}) = 2) = 1$ then each individual always gives birth to exactly 2 offspring—a binary splitting process*.*

The most important example for the purposes of this dissertation is the Yule process:

Example 2.1.8. *(Yule processes[4])*

A process starting with one individual where all individuals live forever and give birth at a unit per-capita rate is a rate-1 Yule process *or* pure birth process*.*

[4]The Yule process dates back to 1925 [114] when it was first used to describe the distribution of the number of species per genus, which Yule observed followed a power law distribution.

That is, if for all individuals x, $\lambda_x = \infty$ and ξ_x is a rate-1 Poisson process, then the branching process is a rate-1 Yule process. If ξ_x is a rate-α Poisson process, then the branching process is a rate-α Yule process.

To describe the extinction and growth rate of such branching processes, it will be useful to distill the reproduction point process ξ_x down into a function:

Definition 2.1.9. *The* reproduction function *of a branching process driven by the life-length and reproduction process (λ, ξ) is*

$$\mu(t) = \mathbb{E}\,\xi([0,t]]$$

so that $\mu(0)$ represents the expected number of offspring born instantly, and $\mu(\infty)$ represents the expected number of offspring born over an individual's entire lifetime.

Since $\mu(\infty)$ in the continuous-time case is the analogue of μ, the expected number of offspring in the discrete-time case, it's unsurprising that continuous-time branching processes follow a similar criticality classification as in the discrete case according to $\mu(\infty)$.

We say a continuous-time branching process is *subcritical, critical,* or *supercritical* according to whether $\mu(\infty) < 1, = 1, or > 1$ respectively. As it turns out, the growth behavior of such processes is still roughly the same as in the discrete case. To state the results rigorously, we need just a bit more notation:

$$z(t) := \text{the number of individuals alive at time } t$$
$$z^a(t) := \text{the number of individuals alive at time } t \text{ who are younger than } a$$

Obviously, if $a > t$ then $z^a(t) = z(t)$.

Things are a bit more complex in the asymptotic analysis of the continuous-time case, so for now we shall content ourselves with showing how to understand the *mean* growth behavior of these processes in terms of $\mathbb{E}\,z(t)$ and $\mathbb{E}\,z^a(t)$.

Recall that in the case of GWBPs (under appropriate moment conditions), the martingale $M_n = Z_n/\mu^n$ tells us that,

1. When $\mu < 1$, $\mathbb{E}\, Z_n \to 0$
2. When $\mu = 1$, $\mathbb{E}\, Z_n \to 1$
3. When $\mu > 1$, $\mathbb{E}\, Z_n = \mu^n$

The same trichotomy persists in continuous time. However, the precise rate of growth in the supercritical, continuous case will depend not only on the total number of offspring $\mu(\infty)$ this time, but on the whole timeline of births μ through a special parameter α:

Definition 2.1.10. *If it exists*[5]*, the* Malthusian rate of growth *of a continuous-time branching process with reproduction function μ is the unique solution α to the equation*

$$\int_0^\infty e^{-\alpha t}\mu(dt) = \int_0^\infty \alpha e^{-\alpha t}\mu(t)dt = 1$$

Where the first equality follows from Fubini's theorem. α is positive, zero, or negative depending on whether $\mu(\infty)$ is $< 1, = 1,$ or > 1.

One can think of the integrand $e^{-\alpha t}$ as the continuous-time analogue of the normalizing sequence μ^n from the supercritical GWBP. And we see that it plays a similar role in the asymptotic analysis.

Theorem 6.3.3 from [65]. *Under some reasonable conditions on μ,*

1. *If $\mu(\infty) < 1$, then*
$$\mathbb{E}\, z(t) \to 0 \text{ as } t \to \infty$$
2. *If $\mu(\infty) = 1$, then for all $0 \leqslant a < \infty$,*
$$\mathbb{E}\, z^a(t) \to C_1 \text{ as } t \to \infty$$

[5]In certain edge cases it does not, for example if $\mu(0) > 1$ then we may have $\mathbb{P}(z(t) = \infty) > 0$.

where C_1 depends only on a, λ, and ξ.

3. *If $\mu(\infty) > 1$, then for all $0 \leqslant a < \infty$,*

$$\mathbb{E}\, z^a(t) \sim e^{\alpha t} C_2$$

where α is the Malthusian rate of growth and C_2 depends only on α, a, λ, and ξ.

Let us see by example what this can tell us.

Example 2.1.11. *(Growth of a rate-ν Yule process)*

The rate-ν Yule process's reproduction is driven by a rate-ν Poisson point process so the reproduction function is given by $\mu(t) = \nu t$. Therefore the Malthusian parameter is solved by

$$1 = \int_0^\infty \alpha e^{-\alpha t} \nu t dt = \nu\alpha \int_0^\infty t e^{-\alpha t} dt = \nu\alpha \left(\frac{1}{\alpha^2}\right)$$

Yielding $\alpha = \nu$. Therefore since $\nu > 0$, the Yule process is supercritical and the population size grows roughly at rate $e^{\nu t}$ as $t \to \infty$.

In Chapter 3, we conduct essentially the same analysis, except instead of a Yule process driven by Poisson point processes, we will study a branching process driven by Yule processes viewed as a point process. We will also make use of stronger limit theorems giving us a.s. and L^1 convergence of $z(t)$.

2.2 2.2. Networks

2.2 2.2.1. Scale-free networks

Networks with power law degree distributions (*scale-free networks*) have experienced a surge of popularity in the past 2 decades. By power law degree distribution, we mean that the degree distribution of a given graph is roughly

$$p(k) = Ck^{-\alpha}, \quad \text{some } \alpha > 1$$

for appropriate values of k. In reality, many real-world networks deviate from this prescription due to finite-size effects. For example, it is common for many social networks to exhibit an exponential cutoff at some large k, see for example [35]. For simplicity in what follows we ignore these effects.

The timing of the surge coincides with the fact that technological advances have allowed us to examine the properties of massive networks such as the Internet and citation networks and discover that many of these have power law degree distributions. Indeed the recent resurgence in the study of scale-free networks can be traced back to Barabasi's empirical discovery that the network of the internet has a power law indegree distribution with $\alpha = 2.1 \pm 0.1$ [4]. Since then many other networks have been shown to exhibit power law degree distributions, spanning a range from networks in social science to the humanities. There are too many examples to name here—see [43] for a more exhaustive discussion—but one particular class of networks is worth mentioning for later reference.

It is known that the social network Twitter exhibits much stronger scale-free characteristics than other popular social networks such as Facebook ([71], [105]). On Twitter, the majority of interactions between users are passive in nature—once a user A "follows" another user B, user A will see all content that user B posts onto the network. This is similar to Facebook, except with one crucial difference. On Facebook a user must *request* a connection ("send a friend request to") with another user and wait for that other user to *approve* the friendship connection before they are connected. On Twitter, the vast majority of users can be followed by any user without a need for the request to be approved. This allows for much higher outdegree distributions which appear closer to a true power law, as shown in [71] and [102], among others.

What makes scale-free graphs interesting relative to those with an exponential tail? The main implication of a power law tail in graphs is the prevalence of high-degree *hubs*. These hubs effectively reduce the shortest-path distance on the graph. For example, it is known that the diameter of the giant component of an $ER(n, p)$ graph scale like $\log n$. In [22] it was

shown that preferential attachment graphs with $m \geqslant 2$ have diameter that instead scale like $\log n / \log \log n$. These results have implications for a wide-range of graph problems, from routing [19] to epidemics and information diffusion [86] (also discussed later).

The study of scale-free graphs has only accelerated recently, but the notion itself is actually quite old. The Erdos-Renyi random graph model was introduced in 1959. Just 6 years later in [88], Price noticed that citation networks exhibit a power law degree distribution. A decade or so later in [89], Price posited the so-called *cumulative advantage* mechanism which generates a network with power law degree distribution according to a simple rich-get-richer scheme: new vertices attach to an existing vertex with probability proportional to the degree of the existing vertex. Over a decade later, this was re-discovered by Barabasi in [10] and is now more commonly known as the *preferential attachment* (PA) mechanism.

This model has become one of the standard workhorses in the complex networks community. At this point it is impossible to compile a representative list of references, we will try to give an overview, restricting ourselves as far as possible to papers close in spirit to this project; see [103] where it was introduced in the combinatorics community, [10] for bringing this model to the attention of the networks community, [82],[43] for survey level treatments of a wide array of models, [23] for the first rigorous results on the asymptotic degree distribution, and [36], [21], [93], and [46] and the references therein for more general models and results.

2.2 2.2.2. Preferential attachment

The canonical way of growing scale-free networks is preferential attachment, and this is the model with which we concern ourselves in the first part of this proposal. There are several variants of the preferential attachment model, but all share the same basic mechanism:

Definition 2.2.1. ***(Preferential attachment)***

1. *Start with two nodes with m edges between them.*[6]

2. *At time n add one vertex with m edges to the existing graph in the following way.*

   (a) *Link the first edge to an existing vertex v with probability proportional to some function $f(D_{n-1}(v))$ where $D_n(v)$ is the degree of v at time n.*

   (b) *Update the degrees of all vertices in the graph*

   (c) *Repeat (a) and (b) until all m edges are connected.*

At time n there will be $2+n$ vertices and total degree $2mn$. For the rest of this section we will concern ourselves with the case of trees $m = 1$, both for simplicity of exposition and because the general case can always be reduced to the case of trees.

The variations in the model concern the function $f(\cdot)$. In the original Barabasi-Albert formulation, $f(D_{n-1}(v)) = D_{n-1}(v)$, i.e. the probability of connecting to an existing vertex v is proportional to the degree of v. The simplest generalization of this model is sometimes referred to as *linear preferential attachment* in which

$$f(D_{n-1}(v)) = D_{n-1}(v) + \alpha, \quad \alpha \geqslant -1$$

Since this model encompasses the original (the special case $\alpha = 0$), we will sometimes let *preferential attachment* mean *linear preferential attachment*. Note that as $\alpha \to \infty$, we have the so-called *uniform attachment* scheme in which new edges attach to existing vertices uniformly at random.

The preferential attachment model has become one of the standard workhorses in the complex networks community, based in part on the fact that it exhibits the power law/heavy tailed degree distribution observed in an array of real world systems. As the literature on preferential attachment is large and very broad, we focus on work that is close in spirit to the

[6]Some formulations of the model begin with a single node and m self-loops. In both cases the limiting behavior is the same. The point is to resolve the difficulty that arises when starting with a single node with no edges, in which case the total degree is 0.

work in this thesis. The preferential attachment model was introduced in the combinatorics community in [103] and was brought to the attention of the networks community in [10]. The papers [82] and [43] give survey-level treatments of a wide array of related models, while [23] gives the first rigorous results on the asymptotic degree distribution. More general models and results can be found in [36], [21], [93], [46], and the references therein.

For all its simplicity, the PA mechanism can be viewed in a deeper light through a continuous-time heuristic which turns out to be very useful. Essentially, the PA process is a type of Polya urn process, which can be embedded into a natural continuous-time process related to the Yule process.

A rate-γ Yule process $\{Y_\gamma(t) : t \geqslant 0\}$ is a continuous-time process which starts at time 0 with 1 individual where individuals in the system survive forever and give birth to new individuals independently of other individuals and at rate γ (i.e. the waiting time between births is $\sim \exp(\gamma)$). When $\gamma = 1$ we shall call this point process the *standard Yule process*.

To foreshadow the connection with our work in the second chapter, let us flesh out this point process explicitly. Suppose that $\{e(k)\}_{k\geqslant 1}$ is a sequence of independent exponential random variables with $e(k) \sim \exp(k)$. If we view these as inter-arrival times of a point process $\mathcal{P}_0$ on $\mathbb{R}^+$, i.e.

$$L(m) = e(1) + \ldots + e(m), \quad \mathcal{P}_0 := (L(1), L(2), \ldots)$$

then $\mathcal{P}_0$ is exactly a standard Yule process.

If we initiate two rate-α Yule processes $Y^1_\alpha(t), Y^2_\alpha(t)$ at the same time, then the numbers of individuals in the processes evolves exactly like a 2-color Polya urn starting with one ball of each color. More precisely let $N^i_\alpha(t)$ denote the number of points in $Y^i_\alpha(t)$ at time t and write $\boldsymbol{Y}(t) = (N^1_\alpha(t), N^2_\alpha(t))$. If $\boldsymbol{X}(n) = (X_1(n), X_2(n))$ is the Polya urn process at step n mentioned above, then

$$\boldsymbol{Y}(\tau_n) \stackrel{\mathrm{d}}{=} \boldsymbol{X}(n), \quad \tau_n = \inf\{t \geqslant 0 : N^1_\alpha(t) + N^2_\alpha(t) = n\}$$

To see this, note that at any fixed time t_0 the probability that Y_α^1 increases by 1 before Y_α^2 does is the probability that the minimum of $Y_\alpha^1(t_0)$ iid $\exp(\alpha)$ random variables is less than the minimum of $Y_\alpha^2(t_0)$ iid $\exp(\alpha)$ random variables, which by the properties of the exponential distribution is proportional to $Y_\alpha^1(t_0)/(Y_\alpha^1(t_0) + Y_\alpha^2(t_0))$. This is exactly the probability that a ball of the first color is picked next in a 2-color Polya urn with $Y_\alpha^1(t_0)$ balls of the first color and $Y_\alpha^2(t_0)$ balls of the second color.

To make the connection to the preferential attachment model, recall that we start with two vertices (labelled 1 and 2) linked with an edge and suppose we have $\alpha = 0$. Clearly the model evolves as the number of balls in a 2-color Polya urn, where the colors correspond to the family lines of either vertex 1 or vertex 2. This simple model is of limited value, but a simple modification yields the bedrock of all later analyses.

First, we need to set up a small variation on the standard Yule process. Let $\{E_\alpha(k) : k \geqslant 1\}$ be a sequence of independent exponential random variables as before except now suppose $E_\alpha(k)$ has rate $k + \alpha$ rather than k. Viewing the above as the inter-arrival times of a point process $\mathcal{P}_\alpha$ on $\mathbb{R}_+$ and setting $L_\alpha(m) = E_\alpha(1) + \cdots + E_\alpha(m)$ for $m \geqslant 1$ as before, define the point process

$$\mathcal{P}_\alpha := (L_\alpha(1), L_\alpha(2), \ldots). \tag{2.3}$$

Note that $\mathcal{P}_0$ is exactly a rate-1 (or *standard*) Yule process. For fixed $t \geqslant 0$, write $N_\alpha(t)$ for the number of points in $\mathcal{P}_\alpha$ which fall in the interval $[0, t]$. This process will drive our key branching process:

Definition 2.2.2 (Continuous time branching process)**.** *Fix $\alpha > 0$. We let $\{\mathsf{BP}_\alpha(t) : t \geqslant 0\}$ be a continuous-time branching process driven by the point process $\mathcal{P}_\alpha$ in* (3.4)*. More precisely:*

(a) *At time $t = 0$ we start with one individual called the root ρ which has offspring distribution $\mathcal{P}_\alpha^\rho \stackrel{d}{=} \mathcal{P}_\alpha$. The times of this point process represent times of birth of new offspring.*

(b) *Every new vertex v that is born into the system is given it's own offspring point process $\mathcal{P}_\alpha^v \stackrel{d}{=} \mathcal{P}_\alpha$, independent across vertices.*

For $t \geqslant 0$, we will view $\mathsf{BP}_\alpha(t)$ as a (random) tree representing the genealogical relationships between all individuals in the population present at time t. Now set:

$$\tau_n := \inf\{t : |\mathsf{BP}(t)| = n\}$$

Writing $\mathcal{T}_n$ for the preferential attachment tree grown with parameter α until size n, we have $\mathsf{BP}(\tau_n) \stackrel{\mathrm{d}}{=} \mathcal{T}_n$, viewed as random rooted trees. But more than that, we have that the two *processes* of growing random trees have the same distribution namely

$$\{\mathsf{BP}(\tau_n) : n \geqslant 1\} \stackrel{\mathrm{d}}{=} \{\mathcal{T}_n : n \geqslant 1\}\,.$$

Thus we have extended the simple idea of an urn process embedded in Yule processes to describe, in continuous time, the evolution of the entire preferential attachment tree. This is the fundamental idea behind our entire analysis of the changepoint regime. Later on we shall derive the Malthusian rate of growth for this process on the way to other limit theorems as well.

2.2 2.2.3. Changepoint detection on networks

The general changepoint detection problem has a vast history owing to its obvious importance in applications such as quality control and reliability of industrial processes, in particular quick detection of process failure in production, as well as fields such as signal processing (e.g. biomedical data including neuronal spike data and seismic data), automatic segmentation of signals into stationary segments via identification of change points etc. An exhaustive overview of the classical literature can be found in [11].

By and large, the statistical theory of changepoint detection is extremely well-developed, see e.g. [39, 24, 27, 28, 97, 98, 99]. On the other hand, the majority of "changepoint

detection" on networks is actually the study of detecting anomalies from what is predicted by the model, without a natural temporal aspect. Indeed there has been a significant amount of work on developing techniques to detect anomalous subgraphs and motifs *within* networks, see e.g. [48, 2, 83, 90, 57, 96], for a wide-ranging survey see [29]. This also includes anomalous edge detection via link prediction algorithms [61].

There has been less work done with temporal (time-varying) networks. Much of the work in this area is centered along quantifying anomalies in a sequence of *deterministic* sequence of graphs, see for example [101], [83], [1]. The changepoint problem on a sequence of a graphs (e.g. across time) with a probabilistic model underneath is less explored and we discuss it here.

Classical changepoint detection is essentially focused on detecting changes in parameters of an *independent* (or stationary) sequence. This is where network changepoint problems start to diverge from the classical regime. First of all, network data is often far from independent, especially if the network size is growing as in dynamic models mentioned below. Secondly, the changes we are interested in often go beyond simple shifts in parameters. We may be interested in more complicated concepts such as changes in community structure or more complicated quantities such as the clustering coefficient. All of these present difficulties when appealing to existing theory.

It is convenient to think of generative network models as falling into one of two classes: *static* models or *dynamic* models. In static models the network size is fixed, whereas in dynamic models the generative mechanism directly models the growth of the network over time. The $ER(n, p)$ Erdos-Renyi random graph on n vertices is a static model, whereas the $PA(1, \alpha)$ preferential attachment is a common dynamic model. This distinction carries over to changepoint problems on networks in a natural way.

First, one can try to detect a change in a static network model across a sequence of realizations of that static model. Adapting classical changepoint techniques in this case is often simpler for a variety of reasons, not least of which is that the network size is taken to be

fixed across time. If the networks in the sequence are taken to be independently generated, then adapting classical changepoint concepts is even more straightforward. The simplest way to do this is to convert the graph sequence to a scalar sequence and apply traditional techniques. In [78] the authors consider a sequence of independent $ER(n, p)$ graphs in which, after a certain point, a subset of nodes begins to connect to each other with higher probability than to the other nodes in the graph. The bulk of the paper is devoted to selecting a proper test statistic, but the stopping rule for actual changepoint estimation still comes back to Average Run Length (ARL) theory. This also is the approach followed in [79]. Even in cases when the graph sequence cannot be easily reduced to a scalar sequence, independence across a static sequence is easy to analyze. For example [87] purport to develop an "entirely general" changepoint detection method using the *generalized hierarchical random graph* model, but at its core their approach is simply maximum likelihood on a sequence of independent graphs.

By and large however, it is unreasonable to assume that from time t to $t + 1$ that the entire graph is independently regenerated, especially if the underlying entities represented by the nodes are fixed. Inspired by voting record graphs from the US Congress (i.e. two congressmen are joined with an edge if they voted together on a bill), [113] has proposed a simple $ER(n, p)$ variant with Markov chain dependence for the presence of an edge between a particular pair of vertices, and analyzed the natural changepoint question arising from that chain. These methods depart somewhat from classical changepoint formula, but still confront a relatively simple situation: a sequence of graphs with a fixed number of vertices and underlying generative model, with a well-understood dependence structure across graphs.

One natural question that has yet to be explored, however, is to investigate changes in parameters controlling dynamic network evolution within a single growing network, specifically, preferential attachment. This is the regime we concern ourselves with in our first project below, and in which virtually no work has been done. Adapting classical techniques is very difficult for the simple reason that dynamic network model evolution is a *highly dependent* process. In preferential attachment for example, the placement of a new node

depends on the entire existing structure of the graph, so individual placements are neither independent, stationary, nor ergodic. This calls for a completely different approach than in standard univariate changepoint analysis.

2.3 2.3. Cascades

2.3 2.3.1. What is a cascade?

In general, a *cascade* is a process occurring on a graph which starts at a single node and spreads across vertices in a way such that affected vertices trigger neighboring vertices in some way. In this way, it is distinguished from other interacting models on a graph (Ising model, voter models) where all particles interact with all other particles simultaneously.

But even then, a *cascade* can mean a lot of different things depending on the context. In the economics literature [14], an informational cascade is a process spreading over a group of individuals in which an individual, having observed the actions of those ahead of him, follows the behavior of the preceding individual without regard to his own information. In this sense, a cascade is a sort of herd mentality process. In other contexts, a cascade can refer to a sort of epidemic model where individuals are carriers for a disease and the disease propagates to other individuals according to some mechanism.

To narrow down our discussion to those models relevant to our study, let us first propose a rough classification. Essentially, all cascade models fall into one or more of these three categories, based on how the mechanism for propagation is defined:

1. **Agent-based models:**

   These models specify the propagation mechanism at the individual level, generally by explicitly modelling a decision-making process for whether or not to propagate the cascade. This process can be simple, such as by observing the proportion of neighbors who are active in the cascade, or can be complex, involving game-theoretic considerations.

In most of these cases, the graph is implied or de-emphasized, e.g. [14] or [9], but there are exceptions. In [37] for example, the authors investigate agent-based diffusions on the classic Watts-Strogatz small world network.

2. **Continuous-time, rate-based models:**

   These models specify the *rate* at which a cascade passes from individuals to one another, and encompasses most epidemic models such at the SI/SIR/SIS models. In these models, the basic dynamic is that uninfected neighbors of infected nodes (*susceptible* notes) become infected at a certain rate and then either stay infected forever or recover at certain other rate. The steady-state is then generally studied using a mean-field approach—approximating the random process flowing on the graph using a set of deterministic differential equations.

   These types of processes are well studied—variations on the SI/SIR/SIS paradigm are especially ubiquitous, see [42] or [115] for typical examples. Therefore we will not go further into it here besides to direct the reader to [46, 85, 63] for comprehensive analyses and overviews.

3. **Discrete-time, probability-based models:**

   These models directly specify the individual *probabilities* of transmission between vertices and is the tradition within which we study them here.

It is worth noting that there is some degree of overlap between all three of these categories. Almost any model in one of these three categories can be reduced to or put in terms of another, if one tries hard enough.

For example, some models (e.g. [109]) specify the propagation of a cascade using a threshold-based approach: susceptible vertices on a graph become infected only when some fixed fraction of their neighbors are infected. This can be viewed as an agent-based model with vertices choosing their state (infected, not infected) based on some function of how

popular each state is among their peers. But the model can also be analyzed in a probabilistic light—by recursively calculating probabilities for whether vertices n steps away from the source of the cascade become infected. Other models such as [60] explicitly involve a rate of infection *and* a threshold model.

However, what does generally remain true to each category are the styles of methods used in and results obtained from each one. The model most relevant to us is the *independent cascade model* first elucidated in [54]. We quote from [70]:

> *The cascade starts with an initial set of active vertices and then unfolds in discrete time according to the following rule. When a vertex v becomes active in step t, then it is given a chance to activate each currently inactive neighbor w; it succeeds with a probability $p_{v,w}$—a parameter of the system—independently of the history thus far. (If w has multiple active neighbors, then their attempts are sequenced in an arbitrary order.) If v succeeds, then w becomes an active vertex in step $n+1$; but whether or not v succeeds, it cannot make any further attempts to activate w in subsequent rounds.*

Indeed, this model is so simple that it is sometimes referred to in the literature as the "cascade" model (as opposed to "threshold" models). Its simplicity makes it amenable to analysis and, in fact, many studies involving estimation of *pairwise transmission probabilities* implicitly study this model. See for example [92, 55]. There is also a long chain of research involving *influence maximization* using this model, i.e. discovering which set of vertices to activate initially in order to achieve the largest resulting cascade, see [70] or [30] for two representative papers on the general theory—see [110] for an extension targeting the Twitter network.

However, not many of these analyses are probabilistic. Rather, they are studies in optimization given a fixed, nonrandom graph. If we move to modelling these cascades on a random graph however, it is not hard to believe that an independent cascade model with

constant probabilities (i.e. $p_{v,w} = \text{some } p \text{ for all } v, w \in V$) can be modelled by some binomial-based branching process. Indeed if a graph is large enough, one would expect a branching process approximation to work quite well in modelling the growth of the cascade. In this thesis our setting will be that of graphs with infinite variance, which, under size-biasing, takes us into the realm of branching processes with varying environment and infinite mean.

However, the most attractive feature of this model to us is the fact that the probabilities $p_{v,w}$ can be varied in a simple way. Not only does this reflect one's intuition about reality, but it also opens the door to the possibility of producing a wide array of cascade behaviors through adjustment of only the probabilities. This general scheme is not new. For example, in [112] the authors propose the closely-related *linear influence model* which also models an independent cascade process whereby the transmission probability for a given cascade is a linear function of the "influence" of the nodes which the cascade has passed through previously. We find this model both more complicated and less salient for the social networks we model. The independent cascade paradigm is also tackled in [55], where the inferential question of learning the transmission probabilities at each node of the network is undertaken. In neither of these cases is a rigorous probabilistic analysis carried out.

2.3 2.3.2. The shape of viral cascades

So how do cascades look like in reality? Do any of the cascades models mentioned above actually do a good job of modeling them? Thanks to smartphones and social media, we have many examples to mine.

On many social networks (not only Twitter, e.g. Facebook), the bulk of interactions are not person-to-person but rather a *broadcast* to all of a user's peers. For example in both Twitter and Facebook, most of the time a user posts content to the social network, it is by default seen by all followers of that user. When a follower of the originating user sees the original content, they then face a decision: if they think it is interesting, they might re-broadcast it to *their* followers, or they can ignore it. On Twitter this is called *retweeting*

a tweet to a user's *followers*, and on Facebook this might be called *resharing* content to a user's *friends*.

In various studies using real cascade data, the dominant finding is that real-life cascades don't match classical theory in a multitude of ways. This isn't surprising. Studies abound about how real cascading behavior occurs at different speeds than is predicted by continuous-time models [62],

However, there is one facet which holds particular interest to us: the *shapes* of real cascades don't match what is predicted by simple models. By shape, we mean the width and depth of the subgraphs traced out by cascades.

To impart some intuition, let's use the example of tweets on Twitter. Suppose we randomly sample tweets from the Twitter social network and look at the resulting subgraph that is traced out by users' successive retweets of the original tweet. Modelling this retweeting process via a branching process means that we are interpreting each retweet as an individual in a GWBP, and each retweet's *successive* retweets as an individual's offspring in a GWBP.

In light of this, there is only one possible way for a tweet with a high number of retweets to get that way. If we view the subgraph of the cascade as a tree, then the tree of a popular tweet necessarily has great depth *and* a high number of individuals at each level. However, the main finding of several recent studies is that real cascades are much richer in shape. There are many tweets with long-lived, yet "skinny" retweet networks. Also common are "star" networks—retweet networks which are very large yet also very shallow. These shapes cannot simultaneously be explained away by easy modifications to a GWBP. Indeed, this is exactly the finding of [52] using data (we shall return to this example again in Chapter 5) and of [44] using data from Facebook.

But this phenomenon is not limited to Twitter, or even Facebook. In [108] the authors study the social network Digg and find that, although there exists content informative enough to potentially reach hundreds of people within one hop of the originator, it rarely if ever affects more than even 0.1% of the entire network. [73] study the cascade of e-mail chain

letters and find that many such cascades trace out a very deep but narrow tree pattern, which as mentioned above is impossible by classical models. See [5] for a study using data from the LinkedIn network and [72] for a study using data scraped from various blogs.

To the author, the simplest explanation for this disconnect between theory and reality is that the transmission probability is not constant over the life of the tweet. By-and-large, most existing models used to explain these phenomena employ the assumption of constant transmission probability over time—in continuous models, this takes the form of a constant *rate*. Almost all epidemic models generally fall in this category and only recently have exceptions cropped up, see [69].

There have been some attempts to resolve these sorts of contradictions with simple branching models. By far, the bulk of these efforts have been by physics working on continuous-time epidemic models, see [38, 68, 75, 63] and also some of the references in the previous two paragraphs.

The crux of the latter portion of this thesis is an analysis of a simple cascading process with decreasing probabilities on a scale-free network topology, which is the correct paradigm for social networks [15].

CHAPTER 3

Changepoint detection on preferential attachment

3.1 3.1. Introduction

Motivated by the availability of data on many real world systems, the last few years have witnessed an explosion in both methodological and theoretical development of various complex network models, see e.g. [20, 46, 106, 33, 82, 81, 3, 43]. One sub-field which has been particularly active is temporal or time varying networks. See the recent surveys [18, 59] and the references therein for more.

Consider the simplest version of the classical (offline) change point detection in the context of *iid* data described as follows. Fix two distribution functions F and G (unknown but different) and a parameter $\gamma \in (0, 1)$. Consider a stream of data $\{X_i : 1 \leqslant i \leqslant n\}$ with distribution: for $i \leqslant \lfloor n\gamma \rfloor$ X_i are *iid* with distribution F whilst for $i > \lfloor n\gamma \rfloor$, X_i are *iid* with distribution G (and independent of the initial segment). Based on the observed data, $\{X_i : 1 \leqslant i \leqslant n\}$, the aim is then to estimate the change point γ using estimators that are consistent as the sample size $n \to \infty$.

Our goal in this chapter is to investigate the analogous problem on temporal networks whose evolution is driven by a mechanism depending on a parameter affected by the change point. Because of their heavily dependent nature, this setting presents some interesting challenges. Our investigation proceeds as follows:

(a) We start by proposing a variant of the standard preferential attachment model which incorporates a change point. This conceptually simple model allows for an easy interpretation of the effect of the change point on network dynamics. We rigorously study the effect of this change point on structural properties of the network including the scale-free

or heavy tailed nature of the limiting degree distribution as well as asymptotics for the maximal degrees.

(b) We then propose and study consistency properties of offline estimation procedures to detect the location of this change point from observed data. In particular this allows one to gain insight into the effect of the non-stationary nature of the evolution of the network model on various known heuristics for estimation in the *iid* setting.

3.1 3.1.1. Organization

Both change point detection as well as preferential attachment models have witnessed enormous amount of work over the last few decades. For a fuller discussion of these two fields, their relevance to this project as well as related work, see Sections 2.2.2 and 2.2.3.

We start in Section 3.1.2 by defining the model. In Section 3.1.3 we setup notation required for the main results. Section 3.2 contains our main results, starting with Section 3.2.1 that describes asymptotics for functionals of the networks including the degree distribution as well as maximal degrees as the network size $n \to \infty$. Section 3.2.2 formulates estimators to find the change point and proves their consistency properties. Proofs for asymptotics of network functionals can be found in Section 3.4. Section 3.4.5 develops a functional central limit theorem for a specific functional of the network. Section 3.4.6 then uses this CLT to prove consistency of the proposed estimator.

3.1 3.1.2. Model formulation

Let's recall the linear preferential attachment model from definition 2.2.1.

Start with a single vertex at time $m = 1$ (this vertex will be referred to as the *root* or the original progenitor of the process and denoted by ρ). Fix a parameter $\alpha \geqslant 0$. At each discrete-time point $1 < m \leqslant n$ a new vertex enters the system with a *single edge*[1] which it

[1]Throughout this chapter we will consider the simplest case where the network at each stage is a tree. The methodology can be generalized to the general network setup.

will then connect to a pre-existing vertex. The vertex connects to a pre-existing vertex v with probability proportional to the current degree of v plus α. Let $\mathcal{T}_m$ denote the graph at time m and $\{\mathcal{T}_m : 1 \leqslant m \leqslant n\}$ be the entire graph valued process. Note that since each new vertex has one edge which it uses to connect to the current graph, $\mathcal{T}_m$ for any m is a tree rooted at ρ. Thus for $m > 1$, the degree of every vertex is at least 1. If we regard the existing vertex to which a new vertex attaches at the *parent* of this vertex, then one can view this process as generating a directed tree with edges pointed from parents to children.

Our analysis is based on a continuous-time version of this process for which a slight variant of the above discrete-time process is more natural. For directed rooted trees the degree of every vertex other than the root is 1+ out-degree of the vertex; for the root, the degree and the out-degree coincide. Fix a single vertex at time $m = 1$ and a parameter $\alpha > 0$. The preferential attachment variant considered in this thesis is as follows: at each stage $m > 1$ a new vertex enters the system and connects to a pre-existing vertex $v \in \mathcal{T}_{m-1}$ with probability proportional to $1 + \alpha$+ out-degree of v in $\mathcal{T}_{m-1}$. This model differs from the original only in the attachment probability to the root, and has all the same asymptotic properties as the original model but is slightly easier to deal with rigorously.

This model has been studied extensively and in particular it is known [23] that the degree distribution converges in the large network limit. Precisely, for fixed $k \geqslant 1$, let $N_n(k)$ denote the number of vertices with degree k in $\mathcal{T}_n$. Then,

$$\frac{N_n(k)}{n} \xrightarrow{\text{a.s.}} p_\alpha(k), \qquad \text{where } p_\alpha(k) := (2+\alpha)\frac{\prod_{j=1}^{k-1}(j+\alpha)}{\prod_{j=3}^{k+2}(j+2\alpha)}. \tag{3.1}$$

Here for $k = 1$, we use the notation $\prod_{j=1}^{k-1} = 1$. Write D_α for a random variable with the above distribution. It is easy to check that there exists a constant $c > 0$ such that

$$\mathbb{P}(D_\alpha \geqslant k) \sim \frac{c}{k^{\alpha+2}}, \qquad \text{as } k \to \infty. \tag{3.2}$$

Further, arranging the degrees in $\mathcal{T}_n$ in decreasing order as $M_n(1) \geqslant M_n(2) \geqslant \cdots M_n(n)$, it is known [80, 12] that for any fixed $k \geqslant 1$, there exists a non-degenerate probability distribution ν_k^α on $\mathbb{R}_+^k$ such that

$$\left(\frac{M_n(j)}{n^{\frac{(1+\alpha)}{(2+\alpha)}}} : 1 \leqslant j \leqslant k \right) \xrightarrow{\text{w}} \nu_\alpha^k. \tag{3.3}$$

where $\xrightarrow{\text{w}}$ denotes weak convergence.

3.1 3.1.2.1. Model with change point

Now fix two attachment parameters $\alpha, \beta > 0$, a change point parameter $\gamma \in (0, 1)$, and a system size $n > 1$. The model does preferential attachment as before, but now the attachment dynamics changes after time $\lfloor n\gamma \rfloor$ namely

(a) For time $0 < m \leqslant \lfloor n\gamma \rfloor$, the new vertex entering the system at time m connects to pre-existing vertices with probability proportional to their current out-degree $+1 + \alpha$.

(b) For time $\lfloor n\gamma \rfloor < t \leqslant n$, the new vertex connects to pre-existing vertices with probability proportional to their current out-degree $+1 + \beta$.

Let $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$ be the driving set of parameters of the model. We will let $\mathcal{T}_{\boldsymbol{\theta},m}$ denote the rooted tree at time m and $\{\mathcal{T}_{\boldsymbol{\theta},m} : 1 \leqslant m \leqslant n\}$ for the entire graph valued process. When the context is clear, for ease of notation we suppress the dependence on $\boldsymbol{\theta}$ and write $\{\mathcal{T}_m : 1 \leqslant m \leqslant n\}$. This model is the main object of interest for the rest of the chapter.

3.1 3.1.3. Preliminary notation

To state our main results we will need to define some additional objects. Recall the parameter set $\boldsymbol{\theta} := (\alpha, \beta, \gamma)$ used to construct the model. Let $\{E_\alpha(k) : k \geqslant 1\}$ be a sequence of independent exponential random variables such that for each fixed $k \geqslant 1$, $E_\alpha(k)$ has rate $k + \alpha$. View the above as the inter-arrival times of a point process $\mathcal{P}_\alpha$ on $\mathbb{R}_+$. More precisely

write,

$$L_\alpha(m) = E_\alpha(1) + \cdots + E_\alpha(m), \qquad m \geqslant 1.$$

Consider the point process

$$\mathcal{P}_\alpha := (L_\alpha(1), L_\alpha(2), \ldots). \tag{3.4}$$

Analogously define $\{E_\beta(k) : k \geqslant 1\}$, $\{L_\beta(k) : k \geqslant 1\}$ and the corresponding point process $\mathcal{P}_\beta$. For fixed $t \geqslant 0$, write $N_\alpha(t) := \mathcal{P}_\alpha[0,t]$ for the number of points in $\mathcal{P}_\alpha$ which fall in the interval $[0,t]$.

We will need variants of the above point process. Fix $j \geqslant 1$ and $\alpha > 0$. Let $\mathcal{P}_\alpha^j$ be the point process where we use the sequence of points $\{E_\alpha(m) : m \geqslant j\}$ to construct the point process so that the first point arrives after an exponential rate $j + \alpha$ amount of time, the second point arrives at rate $j + 1 + \alpha$ after the first point and so forth. As before let $N_\alpha^j(\cdot)$ be the corresponding counting process and note that $N_\alpha^1(\cdot) = N_\alpha(\cdot)$.

Define the constant

$$a = \frac{1}{2+\beta} \log \frac{1}{\gamma}. \tag{3.5}$$

On the interval $[0,a]$, define the "truncated" exponential distribution described via the cumulative distribution function

$$G_a(s) = \frac{1 - \exp(-(2+\beta)s)}{1 - \exp(-(2+\beta)a)}, \qquad s \in [0,a]. \tag{3.6}$$

Write Age for a random variable with distribution G_a (the reason for this terminology will become clear in the proof). Generate a counting process $N_\beta(\cdot)$ as above (independent of Age) so that $N_\beta[0, \mathsf{Age}]$ is the number of points that occur before the random time Age.

We are now in a position to define the limiting degree distribution. Consider the following integer valued random variable $D_{\boldsymbol{\theta}}$:

(a) With probability $1 - \gamma$, $D_{\boldsymbol{\theta}} = 1 + N_\beta[0, \mathsf{Age}]$

(b) With probability γ, $D_{\boldsymbol{\theta}} = D_\alpha + N_\beta^{D_\alpha}[0,a]$ where D_α is a random variable with distribution as in (3.1), namely the limiting degree distribution *without* change point. More precisely, generate D_α with distribution as in (3.1). Conditional on D_α, generate the point process $N_\beta^{D_\alpha}$ and count the number of points in the interval $[0,a]$ and add this to the original random variable D_α.

Write $\mathbf{p}_{\boldsymbol{\theta}} = (p_{\boldsymbol{\theta}}(k) : k \geqslant 1)$ for the probability mass function of the above random variable namely

$$p_{\boldsymbol{\theta}}(k) = \mathbb{P}(D_{\boldsymbol{\theta}} = k), \qquad k \geqslant 1. \tag{3.7}$$

3.2 3.2. Results

Let us now describe our main results. We state results about the asymptotic degree distribution in Section 3.2.1. We formulate statistical procedures to estimate the change point and the associated consistency results in Section 3.2.2.

3.2 3.2.1. Asymptotics for the degree distribution

Fix $\boldsymbol{\theta} \in \mathbb{R}_+ \times \mathbb{R}_+ \times (0,1)$. For fixed $k \geqslant 1$ let $N_n(k)$ denote the number of vertices with degree k in the random tree $\mathcal{T}_n$ constructed in the change point model as in Section 3.1.2.1. The random variable $D_{\boldsymbol{\theta}}$ in the following result is as defined in (3.7).

Theorem 3.2.1. *Fix $k \geqslant 1$. As $n \to \infty$ the degree distribution satisfies,*

$$\frac{N_n(k)}{n} \xrightarrow{\mathrm{P}} \mathbb{P}(D_{\boldsymbol{\theta}} = k), \quad \text{as } n \to \infty$$

Further for $\alpha \neq \beta$ and $\gamma \in (0,1)$, $\mathbf{p}_{\boldsymbol{\theta}} \neq \mathbf{p}_\alpha$. However there exist constants $0 < c < c'$ such that for all $k \geqslant 1$

$$\frac{c}{k^{\alpha+2}} \leqslant \mathbb{P}(D_{\boldsymbol{\theta}} \geqslant k) \leqslant \frac{c'}{k^{\alpha+2}}. \tag{3.8}$$

Remark 1. This theorem says that one **does feel** the effect of the change point in the empirical degree distribution if $\alpha \neq \beta$ and $\gamma \in (0,1)$, however comparing (3.8) with (3.2), for any fixed $\gamma \in (0,1)$, this does **not** change the tail behavior. This is a little surprising as one might assume, especially for γ close to zero and $\beta < \alpha$ (where the no change point dynamics with β instead of α results in a degree distribution with a heavier tail), the tail of the degree distribution might scale like $k^{-(2+\beta)}$, namely the dynamics of attachment driven by β should kick in. However this is not the case.

Remark 2. The techniques developed in this chapter easily extend to the setting of multiple change points. We describe these extensions in Theorem 3.3.1.

The next result deals with maximal degree asymptotics. As before arrange the degrees in $\mathcal{T}_n$ in decreasing order as $M_n(1) \geqslant M_n(2) \geqslant \cdots M_n(n)$.

Theorem 3.2.2. *Fix $k \geqslant 1$ and consider the k maximal degrees $(M_n(j) : 1 \leqslant j \leqslant k)$. Then the sequence of $\mathbb{R}_+^k$ valued random variables defined by setting*

$$\mathbb{M}_n(k) := \left(\frac{M_n(j)}{n^{\frac{(1+\alpha)}{(2+\alpha)}}} : 1 \leqslant j \leqslant k \right), \qquad n \geqslant 1,$$

is tight and bounded away from zero.

Remark 3. Comparing the scaling of the maximal degrees above to the setting of no change point as described in (3.3), one sees that the maximal degrees do not feel the effect of the change point, at least in terms of their order of magnitude. We further conjecture that $\{\mathbb{M}_n(k) : n \geqslant 1\}$ converge weakly to a non-degenerate distribution on $\mathbb{R}_+^k$. We have not pursued this further.

3.2 3.2.2. Change point detection

In this section we formulate a non-parametric estimator for the change point based on observations of the network and establish its consistency. While one could use the explicit

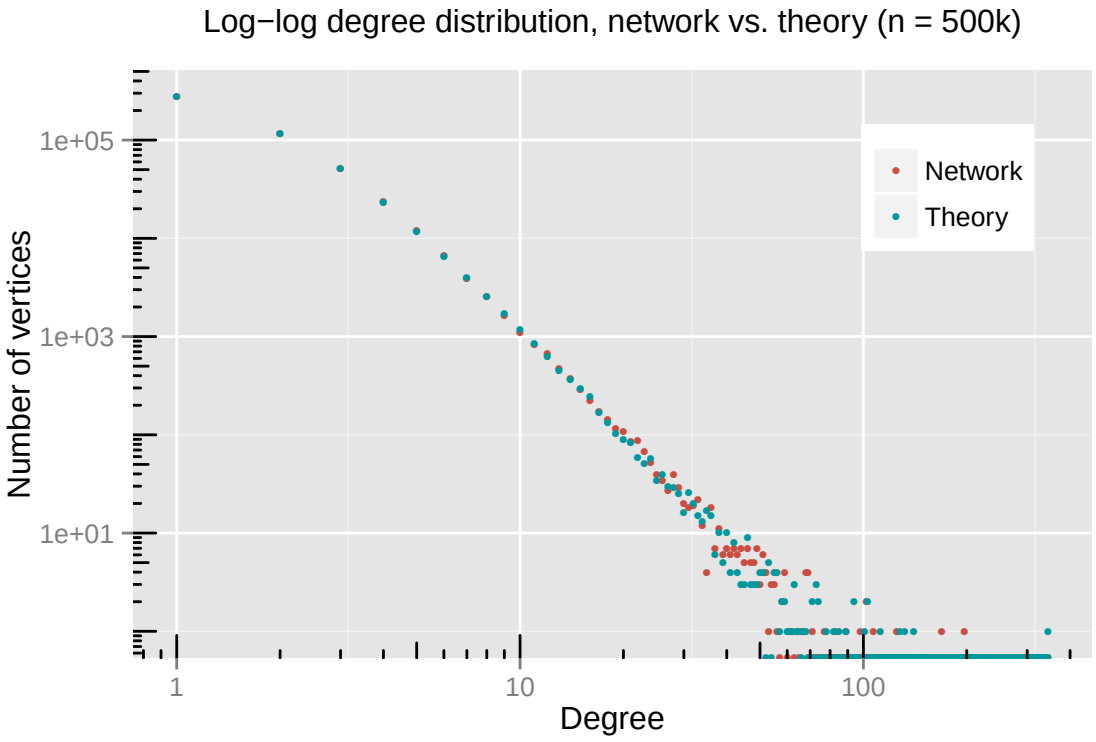


Figure 3.1: Log-log plot of the limiting degree distribution (red) and simulated network degree distribution (blue) with network size $n = 500,000$ and a corresponding sample of the same size from the predicted degree distribution. The model parameters are taken as $\alpha = 6, \beta = 1$ and the change point $\gamma = 0.5$.

linear nature of the attachment scheme to devise parametric or likelihood-based estimators of the change point, our aim is to develop more flexible methods that may work in settings where the precise form of the attachment model before and after the change point is not known. The plan of the rest of this section is as follows. Our estimator tracks the proportion of leaves as the process evolves and uses this functional to formulate a non-parametric estimator. Thus we start by describing a functional central limit for the proportion of leaves (Theorem 3.2.3). Then we formulate the actual estimator based on this functional. Theorem 3.2.3 is then used to establish the consistency result (Theorem 3.2.4) for the proposed estimator.

For fixed $k \geqslant 1$ let $N_n(k, m)$ denote the number of vertices with degree k in the tree $\mathcal{T}_n$ *at the time of appearance of the* m*th vertex*. Rescaling time by n, for $0 \leqslant t \leqslant 1$, let $\hat{N}_n(k, t) = N_n(k, nt)$ and let

$$\hat{p}^n(k, t) = \frac{\hat{N}_n(k, t)}{nt}, \qquad 0 \leqslant t \leqslant 1, \tag{3.9}$$

be the proportion of vertices with degree k at time nt. The $k = 1$ case corresponds to the number of leaves. To ease notation in the displays below, write $\hat{p}^n(1, t) = \hat{p}^n_t$. Now define the continuous function,

$$p_t^{(\infty)} = \begin{cases} \frac{2+\alpha}{3+2\alpha} & \text{if } 0 \leqslant t \leqslant \gamma \\ \frac{2+\beta}{3+2\beta}\left(1 - \left(\frac{\gamma}{t}\right)^{\frac{3+2\beta}{2+\beta}}\right) + \frac{\gamma}{t}\left(\frac{2+\alpha}{3+2\alpha}\right)\left(\frac{\gamma}{t}\right)^{\frac{1+\beta}{2+\beta}} & \text{if } \gamma \leqslant t \leqslant 1. \end{cases} \tag{3.10}$$

We will prove in Section 3.4.5.1 that for each fixed $0 < t \leqslant 1$, $p_t^{(\infty)}$ will represent the limiting proportion of leaves in $\mathcal{T}_{nt}$. To simplify notation in the sequel, define the function $\delta : \mathbb{R}_+ \to [0, 1]$ by the prescription

$$\delta_u := \frac{1+u}{2+u}, \qquad u \geqslant 0. \tag{3.11}$$

Note that $p_t^{(\infty)} = p_\gamma^{(\infty)}$ for $t \leqslant \gamma$. Now define the positive function $\{\sigma_M(t) : 0 \leqslant t \leqslant 1\}$ via the formulae

$$\sigma_M^2(t) := \begin{cases} t^{2\delta_\alpha} \left[\delta_\alpha p_\gamma^{(\infty)}(1 - \delta_\alpha p_\gamma^{(\infty)})\right] & \text{if } 0 \leqslant t \leqslant \gamma, \\ \\ \gamma^{2\delta_\alpha} \left(\frac{t}{\gamma}\right)^{2\delta_\beta} \delta_\beta p_t^{(\infty)}(1 - \delta_\beta p_t^{(\infty)}), & \text{if } \gamma < t \leqslant 1. \end{cases} \tag{3.12}$$

For later use define the functions

$$\sigma^2(t) := \begin{cases} \left[\delta_\alpha p_\gamma^{(\infty)}(1 - \delta_\alpha p_\gamma^{(\infty)})\right] & \text{if } 0 \leqslant t \leqslant \gamma, \\ \\ \delta_\beta p_t^{(\infty)}(1 - \delta_\beta p_t^{(\infty)}), & \text{if } \gamma < t \leqslant 1, \end{cases} \tag{3.13}$$

and

$$\mu(t) := \begin{cases} -\frac{\delta_\alpha}{t^{\delta_\alpha + 1}} & 0 < t \leqslant \gamma \\ \\ -\frac{\delta_\beta \gamma^{\delta_\beta - \delta_\alpha}}{t^{\delta_\beta + 1}} & \gamma < t \leqslant 1 \end{cases} \tag{3.14}$$

Define the diffusion $\{M(t) : 0 \leqslant t \leqslant 1\}$ via the prescription

$$dM(t) = \sigma_M(t) dB(t), \qquad 0 \leqslant t \leqslant 1. \tag{3.15}$$

Here $\{B(u) : u \geqslant 0\}$ is standard Brownian motion on $\mathbb{R}_+$. Thus M is essentially a deterministic time change of $B(\cdot)$ namely

$$\phi(t) = \int_0^t \sigma_M^2(s) ds, \qquad \{M(t) : 0 \leqslant t \leqslant 1\} \stackrel{d}{=} \{B(\phi(t)) : 0 \leqslant t \leqslant 1\}. \tag{3.16}$$

In particular $M(\cdot)$ is a Gaussian process on $[0,1]$. Finally define the functions

$$g(t) := \begin{cases} \frac{1}{t^{\delta\alpha}} & \text{if } 0 < t \leqslant \gamma, \\ \\ \frac{\gamma^{\delta\beta-\delta\alpha}}{t^{\delta\beta}} & \text{if } \gamma < t \leqslant 1. \end{cases} \tag{3.17}$$

Define the process

$$G(t) = g(t)M(t), \qquad 0 < t \leqslant 1. \tag{3.18}$$

By Ito's formula $G(\cdot)$ solves the SDE

$$dG(t) = \mu(t)M(t)dt + \sigma(t)dB(t), \tag{3.19}$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ are as in (3.13) and (3.14) respectively. Then we have the following result.

Theorem 3.2.3. *Consider the process of re-centered and normalized number of leaves*

$$G_n(t) := \frac{\hat{N}_n(1,t) - ntp_t^{(\infty)}}{\sqrt{n}}, \qquad 0 \leqslant t \leqslant 1, \tag{3.20}$$

with linear interpolation between time points. Then as $n \to \infty$, $G_n \xrightarrow{\text{w}} G$ where G is the diffusion defined in (3.19) *and $\xrightarrow{\text{w}}$ denotes weak convergence on $D[0,1]$ equipped with the usual Skorohod metric.*

For the rest of this section, let $p_n(m)$ denote the proportion of leaves (degree one vertices) in $\mathcal{T}_m$. Fix $\varepsilon > 0$. We will define two functions on the interval $[\varepsilon, 1]$. Let

$${}_th^{(n)} = \frac{1}{n(t-\varepsilon)} \sum_{m=n\varepsilon}^{nt} p_n(m), \qquad \varepsilon \leqslant t \leqslant 1. \tag{3.21}$$

Let

$$h_t^{(n)} = \frac{1}{n(1-t)} \sum_{m=nt+1}^{n} p_n(m), \qquad \varepsilon \leqslant t \leqslant 1. \tag{3.22}$$

In words, ${}_th^{(n)}$ represents the average proportion of leaves in the process between time $n\varepsilon$ and nt while $h_t^{(n)}$ represents the same quantity but after time nt. Define the function

$$D_n(t) := (1-t)|{}_th^{(n)} - h_t^{(n)}|, \qquad t \in [\varepsilon, 1]. \tag{3.23}$$

Write $\mathcal{M}_n$ for the collection of points t for which the corresponding function value $D_n(t)$ is within $\log n/\sqrt{n}$ of the maximum of the function. Precisely, let $D_n^* = \max_{t\in[\varepsilon,1]} D_n(t)$ and let

$$\mathcal{M}_n := \left\{ t \in [\varepsilon, 1] : |D_n(t) - D_n^*| \leqslant \frac{\log n}{\sqrt{n}} \right\}. \tag{3.24}$$

Finally let

$$\hat{\gamma}_n := \max\left\{ t : t \in \mathcal{M}_n \right\}. \tag{3.25}$$

The functionals $D_n^*, \mathcal{M}_n,$ and $\hat{\gamma}_n$ all depend on ε but we suppress this dependence to ease exposition below.

Theorem 3.2.4. *Assume that the change point $\gamma > \varepsilon$. Then the estimator $\hat{\gamma}_n \xrightarrow{\text{P}} \gamma$ and in fact*

$$|\hat{\gamma}_n - \gamma| = O_P\left(\frac{\log n}{\sqrt{n}}\right) \tag{3.26}$$

Thus $\hat{\gamma}_n$ is a consistent estimator for the change point γ.

Remark 4. The ε-truncation away from zero is a technical compensation for the factor t in the denominator in (3.21). Technically one should be able to choose a sequence $\varepsilon_n \downarrow 0$ slowly enough such that the above result (modified using this sequence ε_n instead of the fixed ε) is true. This would make the assumption of $\gamma > \varepsilon$ irrelevant in the statement of the Theorem.

Remark 5. The threshold $\log n/\sqrt{n}$ in (3.24) was arbitrary in the sense that if we chose the threshold to be $\omega_n/\sqrt{n}$ where $\omega_n \to \infty$ arbitrarily slowly then the corresponding estimator would satisfy (3.26) with bound $\omega_n/\sqrt{n}$.

Remark 6. See Figure 3.2 for a figure based on simulations for the function $D_n(t)$ with ε taken to be 0.025.

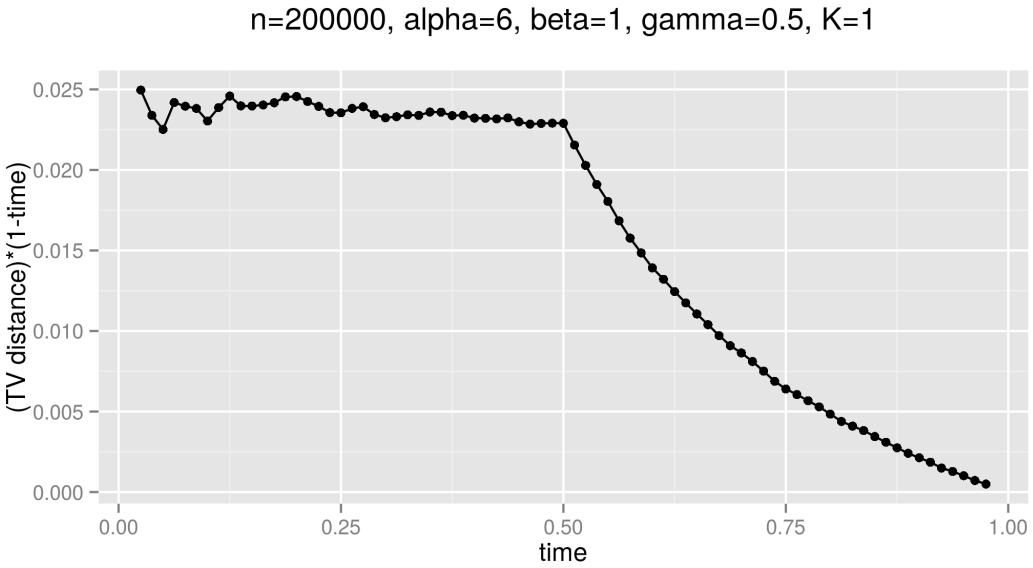


Figure 3.2: The function $D_n(t)$ with network size $n = 200,000$, and model parameters $\alpha = 6, \beta = 1$ and the change point $\gamma = .5$ as in Figure 3.1.

3.3 3.3. Discussion

We now discuss the relevance of our results, their connections to existing literature and possible extensions of our results.

3.3 3.3.1. Change point detection literature

As already mentioned in Chapter **??**, this problem has a vast history owing to its obvious importance in many fields. We refer the reader back to that chapter for more details. Here, we spend just a little time elucidating the math formulation of the classical change point problem

In this context, recall the motivating example of an independent stream of data $\{X_i : 1 \leqslant i \leqslant n\}$ with a change point in the distribution from F to G at time $n\gamma$ described

in Section 3.1. Let ${}_tH^{(n)}(\cdot)$ and $H_t^{(n)}$ denote the empirical distribution of the data before and after t namely

$$ {}_tH^{(n)} := \frac{1}{nt}\sum_{i=1}^{nt}\delta_{X_i}, \qquad H_t^{(n)} := \frac{1}{n(1-t)}\sum_{i=nt+1}^{n}\delta_{X_i}, \qquad 0<t<1. $$

Now define

$$ D_n(t) := t^{1/2}(1-t)^{1/2}\,\mathrm{dist}({}_tH^{(n)}, H_t^{(n)}), \tag{3.27} $$

where dist is any standard notion of distance between probability distributions on $\mathbb{R}$ e.g. Kolmogarov-Smirnov supremum norm or total variation distance. Finally define

$$ \hat{\gamma}_n = \arg\max_{t\in[0,1]} D_n(t). $$

Then in [27] it is shown that $\hat{\gamma}_n$ is a consistent estimator of γ. This was partial motivation for our estimator. Note the "asymmetry" as a function of t between the "classical" context and the model with change point highlighting the non-ergodic nature of the evolution of the model after the change point. We will dive a little deeper into this interesting point shortly.

A second point to note is that we use information on leaf densities in the large network $n \to \infty$ limit. As in [91], one should be able to build on the functional CLT for leaf counts to establish a joint functional CLT for $\left\{\hat{N}_n(k,t) : 1 \leqslant k \leqslant K, 0 \leqslant t \leqslant 1\right\}$ after proper normalization and re-centering for any fixed $K \geqslant 1$. Modifying the estimator in Section 3.4.6 should enable one to get estimators that perform better for finite n.

3.3 3.3.2. The asymmetry within the scaling $(1-t)$

As mentioned above, our estimator is heavily inspired by and very reminiscent of the estimator of [27]. So what's different? Let us investigate the question by making an observation.

In all cases, a change in the attachment parameter manifests in a change in the proportion of leaves. So why can't we build a change point estimator based solely on an *unscaled* version

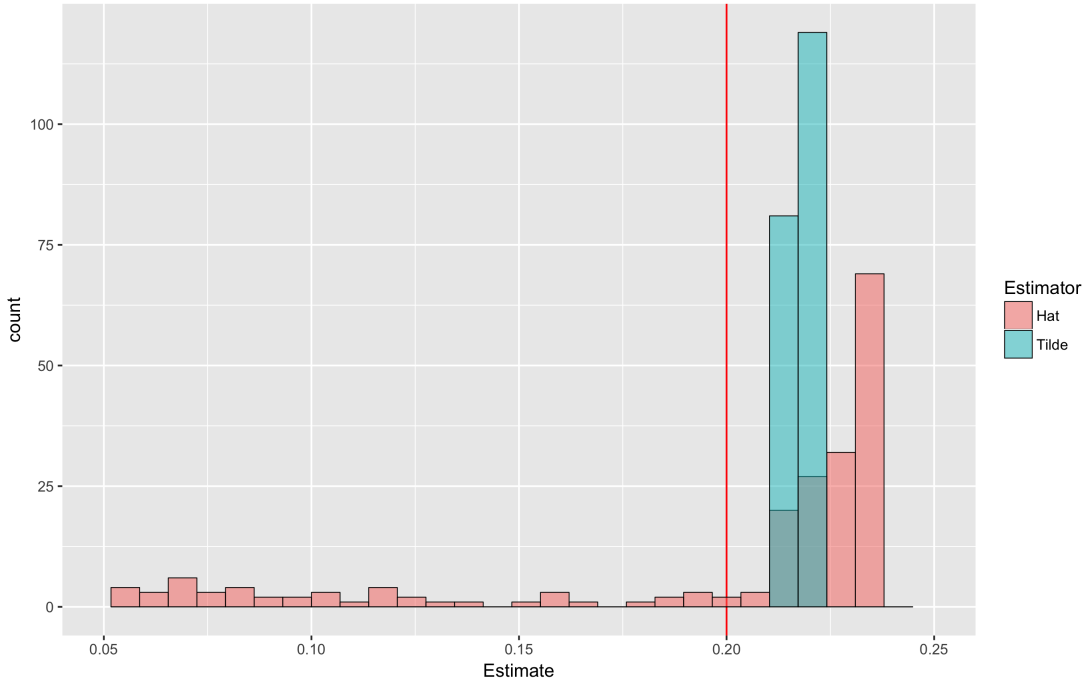


Figure 3.3: Histograms of $\tilde{\gamma}$ vs. $\hat{\gamma}$ for a change point of $\alpha = 0$ to $\beta = 10$ at $\gamma = 0.20$ ($N = 100,000$ vertices).

of D_n? That is, what's stopping us from defining another estimator $\tilde{\gamma}$ by:

$$\tilde{D}_n(t) := |{}_t h^{(n)} - h_t^{(n)}|, \qquad t \in [\varepsilon, 1]. \tag{3.28}$$

$$\tilde{\gamma} = \text{argmax}_{t \in [\varepsilon,1]} \tilde{D}_n(t)$$

The estimator $\tilde{\gamma}$ actually performs quite well for γ close to 0, (Figure 3.3).

However, the same is not true for γ close to 1 (Figure 3.4).

Earlier we alluded to the asymmetry in the change point problem on this temporal network due to strong dependency. This is the crux of the matter—when the attachment parameter changes, the graph doesn't feel the full effect immediately. Rather, it slowly evolves and the effect isn't felt fully until very far away from the changepoint.

Observe in Figure 3.5 that the proportion of leaves begins to change immediately upon hitting the change point at $t = 0.9$, decreasing (almost) linearly. However, because the change point occurs so late and there is heavy dependence in the graph, the proportion of

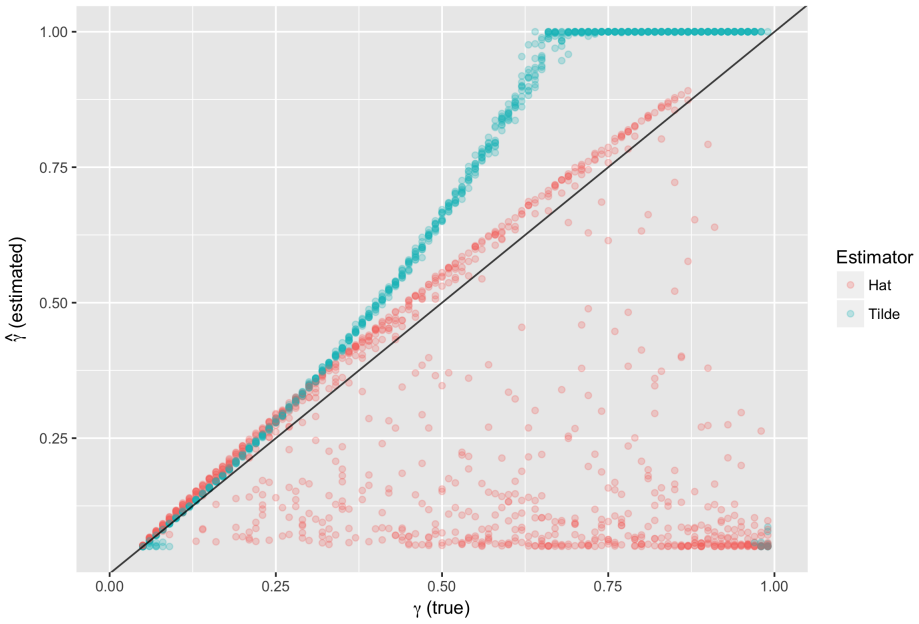


Figure 3.4: $\tilde{\gamma}$ vs. $\hat{\gamma}$ for a change point of $\alpha = 0$ to $\beta = 10$ at various values of γ ($N = 100,000$ vertices).

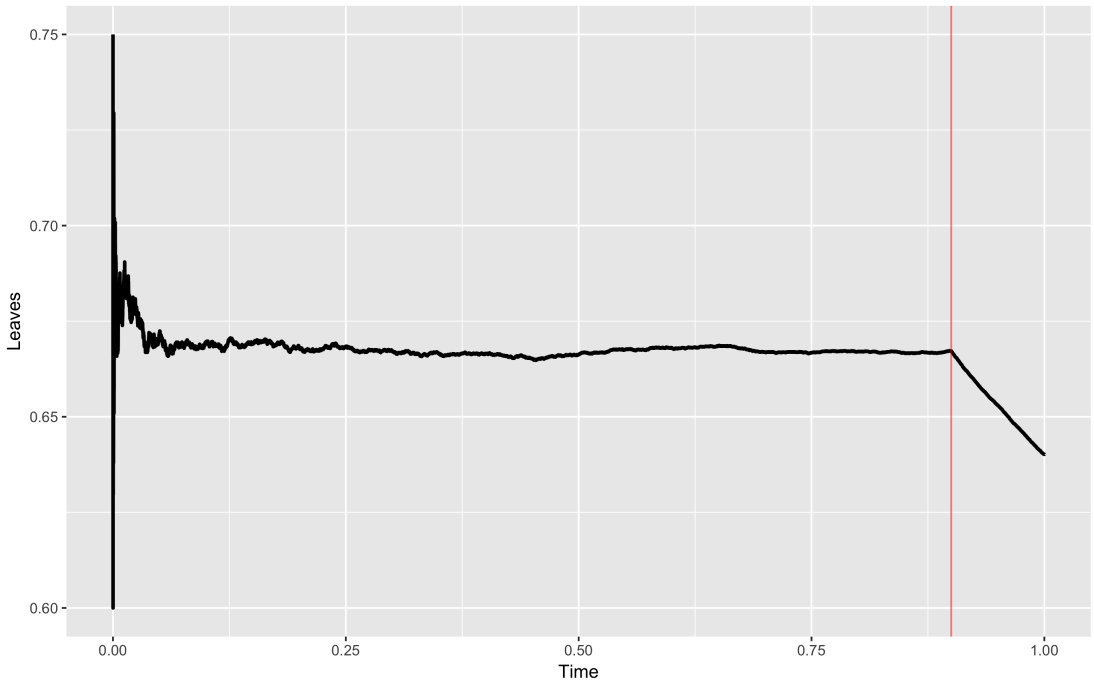


Figure 3.5: The proportion of leaves in a PA graph on $N = 100,000$ vertices with change point at $\gamma = 0.9$ from $\alpha = 0$ to $\beta = 10$.

leaves does not jump to its lower limit proportion immediately—it smoothly decreases and therefore with the unscaled $\tilde{D}_n(t)$, the obvious argmax of this function occurs at $t = 1$. For smaller γ closer to 0.5, say, the effect is less drastic but still apparent.

In general, the basic idea is this: When γ occurs later, the early part of the path of $D_n(t)$ needs to be "discounted" at a certain rate so that the change in the proportion of leaves is felt. For preferential attachment, that rate is $(1 - t)$.

To put it another way, this scaling tells us exactly how different our problem is from the classical one with iid data. One can see this by way of comparison with the setting of [27] of the previous subsection; the analogous $D_n(t)$ function for the iid setting in equation 3.27 carries a scaling of $t^{1/2}(1 - t)^{1/2}$.

3.3 3.3.3. Multiple change points

The proof techniques carry over in a straightforward fashion to the general setting of multiple change points. Fix time points $0 < \gamma_1 < \gamma_2 < \cdots \gamma_k < 1$ and parameters $\alpha, (\beta_i)_{1 \leqslant k}$. As before write $\boldsymbol{\theta} = (\alpha, (\beta_i)_{1 \leqslant i \leqslant k}, (\gamma_i)_{1 \leqslant i \leqslant k})$ for the parameter set. Consider the random tree $\mathcal{T}_n = \mathcal{T}_{\boldsymbol{\theta},n}$ where

(i) In the interval $\{1 < t \leqslant \gamma_1 n\}$, vertices use the attachment scheme driven by α (namely each new vertex attaches to an existing vertex with probability proportional to out-degree $+1 + \alpha$).

(ii) In subsequent intervals $\{\gamma_j n < t \leqslant \gamma_{j+1} n\}$ where $1 \leqslant j \leqslant k$, vertices perform the attachment scheme driven by the parameter β_j. Here we use the convention $\gamma_0 = 0, \gamma_{k+1} = 1$.

As in Section 3.1.3 define the point processes $\mathcal{P}_\alpha, \mathcal{P}_{\beta_i}$ and for fixed $j \geqslant 1$, the point processes $\mathcal{P}_\alpha^j, \mathcal{P}_{\beta_i}^j$. To simplify notation, for any $t \geqslant 0$ and point process $\mathcal{P}$, set $\mathcal{P}[0, t]$ for the *number*

of points in the interval $[0, t]$. Define the constants

$$\pi_j = \gamma_{j+1} - \gamma_j, \qquad a_j = \frac{1}{2+\beta_j} \log \frac{\gamma_{j+1}}{\gamma_j}. \tag{3.29}$$

Note that $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_k)$ is a probability mass function. Write Epoch for a random variable with distribution $\boldsymbol{\pi}$ (i.e. $\mathbb{P}(\mathsf{Epoch} = i) = \pi_i$ for $0 \leqslant i \leqslant k$). Using the constants $\{a_i : 1 \leqslant i \leqslant k\}$ let G_{a_i} denote corresponding truncated exponential distributions as in (3.6) and let Age_i denote a random variable with distribution G_{a_i}. Now construct the random variable $\mathsf{TimeAlive}$ as follows:

(a) Generate a collection of independent random variables Epoch and $\{\mathsf{Age}_i : 1 \leqslant i \leqslant k\}$ with distributions specified as above.

(b) Conditional on $\mathsf{Epoch} = i$, let

$$\mathsf{TimeAlive} = \mathsf{Age}_i + \sum_{j=i+1}^{k} a_j,$$

where again by convention, if $\mathsf{Epoch} = 0$, $\mathsf{Age}_0 = 0$ and so $\mathsf{TimeAlive} = \sum_{j=1}^{k} a_i$.

Construct a positive integer valued random variable $D_{\boldsymbol{\theta}}$ as follows:

(i) Generate $\mathsf{Epoch} \sim \boldsymbol{\pi}$ as above and the corresponding random variable $\mathsf{TimeAlive}$.

(ii) If Epoch takes a non-zero value $1 \leqslant i \leqslant k$, conditional on $\mathsf{Epoch} = i$, generate the switching point process $\mathcal{P}_\star$ on the interval $[0, \mathsf{TimeAlive}]$ as follows:

(a) **Initialization:** In the interval $[0, \mathsf{Age}_i]$, start with $\mathcal{P}_\star = \mathcal{P}_{\beta_i}$. Suppose by time Age_i, $\mathcal{P}_\star[0, \mathsf{Age}_i] = k$. Now generate a point process $\mathcal{P}^{k+1}_{\beta_{i+1}}$ and let $\mathcal{P}_\star[0, \mathsf{Age}_i + a_{i+1}] = \mathcal{P}_\star[0, \mathsf{Age}_i] + \mathcal{P}^k_{\beta_{i+1}}[0, a_{i+1}]$.

(b) **Recursion:** For each subsequent interval $[a_j, a_{j+1}]$ with $j > i$, conditional on $\mathcal{P}_\star[0, \mathsf{Age}_i + a_{i+1} + \cdots a_j] = k_j$, generate the point process $\mathcal{P}_{\beta_{j+1}}^{k_j+1}$. Define

$$\mathcal{P}_\star[0, \mathsf{Age}_i + a_{i+1} + \cdots a_{j+1}] = \mathcal{P}_\star[0, \mathsf{Age}_i + a_{i+1} + \cdots a_j] + \mathcal{P}_{\beta_{j+1}}^{k_j+1}[0, a_{j+1}].$$

Iterate until the last interval resulting in $\mathcal{P}_\star[0, \mathsf{TimeAlive}]$.

Now define $D_{\boldsymbol{\theta}} = 1 + \mathcal{P}_\star[0, \mathsf{TimeAlive}]$.

(iii) If $\mathsf{Epoch} = 0$, so that $\mathsf{TimeAlive} = a_1 + \cdots a_k$, generate a random variable D_α with distribution $\mathbf{p}_\alpha$ as in (3.1). Conditional on D_α, generate $\mathcal{P}_\star$ in the interval $[0, a_1]$ with distribution $\mathcal{P}_{\beta_1}^{D_\alpha}$ and then sequentially proceed as in (ii). In this case, define $D_{\boldsymbol{\theta}} = D_\alpha + \mathcal{P}_\star[0, \mathsf{TimeAlive}]$.

Write $p_{\boldsymbol{\theta}}(\cdot)$ for the pmf of $D_{\boldsymbol{\theta}}$. As before for $k \geqslant 1$, let $N_n(k)$ denote the number of vertices with degree k in $\mathcal{T}_n$. Then we have the following result.

Theorem 3.3.1. *As $n \to \infty$ we have*

$$\frac{N_n(k)}{n} \xrightarrow{\mathrm{P}} p_{\boldsymbol{\theta}}(k).$$

Further there exist constants $0 < c < c'$ such that for all $k \geqslant 1$

$$\frac{c}{k^{\alpha+2}} \leqslant \mathbb{P}(D_{\boldsymbol{\theta}} \geqslant k) \leqslant \frac{c'}{k^{\alpha+2}}. \tag{3.30}$$

3.3 3.3.4. Existing work regarding preferential attachment

We are not aware of other analysis of the effect of change point in structural properties of such network models. However there has been some recent interest in understanding and detecting the "initial seed" [26, 25, 40]. Here one starts with an initial "seed graph" at time $m = 0$ and then performs preferential attachment started from that seed. The aim is then to

estimate this initial seed based on an observation of the network at some large time n. While different from this thesis, this body of work again emphasizes the sensitive dependence on initial conditions for such network models.

3.3 3.3.5. Proof techniques

A number of techniques have been developed to rigorously analyze functionals such as asymptotic degree distributions (see [46, 106] for nice pedagogical treatment). The standard technique involves writing down recursions for the expected degree distribution $\mathbb{E}(N_n(k))$ using the prescribed dynamics of the process, to show that these expectations (normalized by n) converge in the limit and then showing that the deviations $|N_n(k) - \mathbb{E}(N_n(k))|$ are small via concentration inequalities.

In this project, for understanding structural properties we use a different technique, essentially embedding the discrete-time model in a corresponding "continuous-time" branching process $\{\mathsf{BP}^n_{\boldsymbol{\theta}}(t) : t \geqslant 0\}$. This is sometimes known as the Athreya-Karlin embedding of urn processes (see discussion in Section 2.2.2 or in [6]). This explains the various point processes that arise in the description of the limiting degree distribution. While mathematically more involved, this technique gives more insight into the results as it elucidates the natural time scale of the process. In various other settings this technique has resulted in the study of much more general functionals of the process such as the spectral distribution of the adjacency matrix [12] and has been used to derive asymptotic results in "non-local" preferential attachment models [13]. In this project the technique also allows one to intuitively understand why the degree exponent does not change. We advise the reader to come back to the text below after going through the proofs but let us explain the basic intuition here.

In the continuous-time version, the process grows exponentially and in particular takes time $\tau_{\gamma n} \approx \frac{1}{2+\alpha} \log \gamma n + O_P(1)$ to get to size $n\gamma$. At this time there is a change in the evolution where each vertex adopts attachment dynamics driven by the parameter β. However owing to the exponential growth rate, the time for the process to get to size n is $\tau_n \approx \tau_{\gamma n} + a$ where

a is as in (3.5). It turns out that this is not enough time for the dynamics with attachment parameter β to change the degree exponent (since we only have to wait an $O(1)$ extra units of time to get to system size n from γn). These ideas are made mathematically rigorous in Section 3.4.3. For the interested reader, much of the foundational work on continuous-time branching processes relevant for this thesis can be found in [66, 67, 65].

3.3 3.3.6. Empirical dependence of the convergence on parameter values

Recall that the Gaussian process defined in (3.19) underlying the main consistency result Theorem 3.2.4 depends on $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$. One consequence of this dependence is that when the parameter values α and β are close, the change point becomes harder to detect in the sense that larger n is required to get good estimates. This is most easily seen in terms of the fluctuations of the proportion of leaves in the graph.

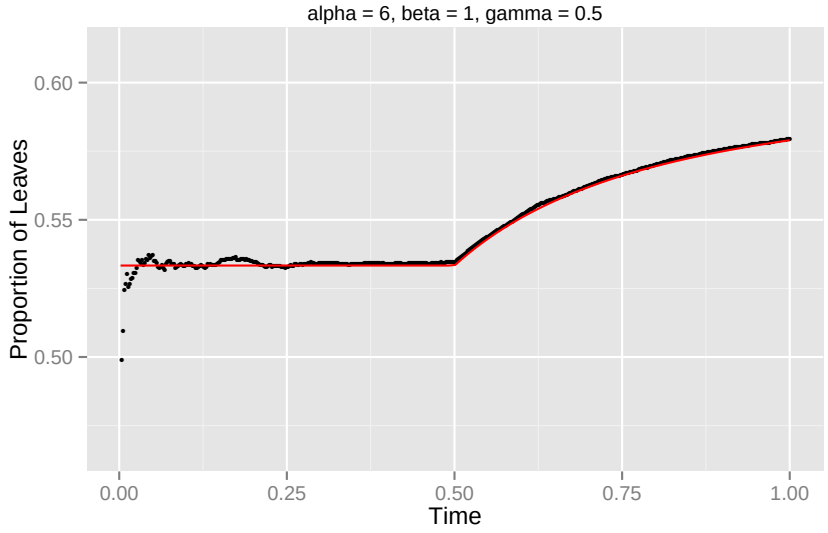


Figure 3.6: Empirical proportion of leaves in a simulation with $n = 200,000, \alpha = 6, \beta = 1, \gamma = 0.5$. The red line represents the theoretical predictions in (3.10).

In both Figures 3.6 and 3.7, the preferential attachment process starts with $\alpha = 6$ and decreases, to $\beta = 1$ in 3.6 and $\beta = 5$ in 3.7. Furthermore the predicted behavior (red line) is almost the same: the proportion of leaves is constant up to the change point $\gamma = 0.5$ and then increases, consistent with a *decrease* in the attachment parameter.

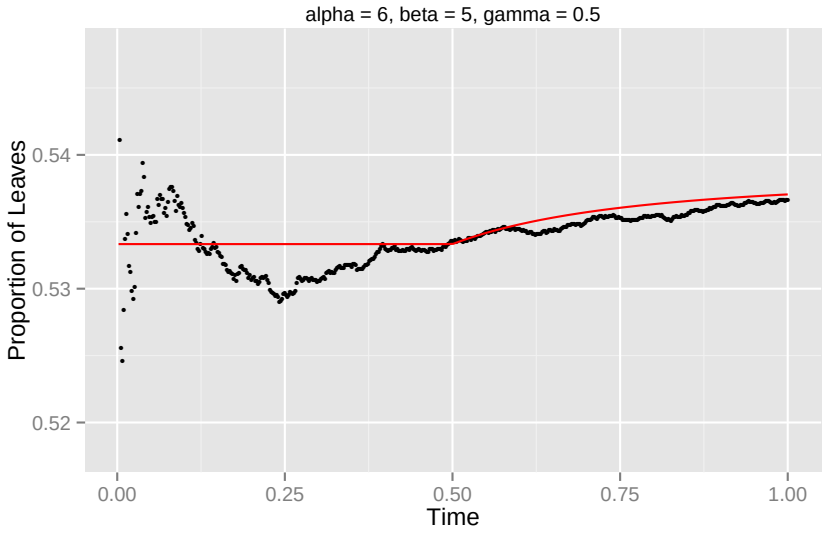


Figure 3.7: Empirical proportion of leaves in a simulation with $n = 200,000, \alpha = 6, \beta = 5, \gamma = 0.5$. The red line represents the theoretical predictions in (3.10).

Despite the sizes of the final graphs in both simulations being $n = 200,000$ vertices, at first glance the fluctuations appear much greater in the latter case. On closer examination however, this is simply an illusion of the axes. In essence, when the shift in parameters is smaller, the change in the proportion of leaves pre- and post-γ is smaller compared to the natural fluctuations in the proportion of leaves which is of order $\sqrt{n}$ (Theorem 3.2.3). Therefore any difference is more difficult to detect for same n. This is not surprising, but worth noting in practice.

3.4 3.4. Proofs

As described in Section 3.3.5, the main conceptual idea is a continuous-time embedding of the discrete-time process. We start in Section 3.4.1 by describing this embedding and deriving simple properties. Then in Section 3.4.2 we prove Theorem 3.2.1. Section 3.4.3 proves the assertion that the degree exponent does not change. Section 3.4.4 analyzes asymptotics for the maximal degrees. Section 3.4.5 contains an in-depth analysis of the density of leaves and

proves Theorem 3.2.3. Section 3.4.6 then uses this theorem to prove the consistency of the estimator namely Theorem 3.2.4.

3.4 3.4.1. Preliminaries

We start with the following definition. To ease notation, for the rest of the chapter we use γn instead of $\lfloor \gamma n \rfloor$.

Definition 3.4.1 (Continuous time branching process)**.** *Fix $\alpha > 0$. We let $\{\mathsf{BP}_\alpha(t) : t \geqslant 0\}$ be a continuous-time branching process driven by the point process $\mathcal{P}_\alpha$ defined in* (3.4)*. Precisely:*

(a) *At time $t = 0$ we start with one individual called the root ρ with an offspring point process with distribution $\mathcal{P}_\alpha^\rho \stackrel{d}{=} \mathcal{P}_\alpha$. The times of this point process represent times of birth of new offspring of ρ.*

(b) *Every new vertex v that is born into the system is given its own offspring point process $\mathcal{P}_\alpha^v \stackrel{d}{=} \mathcal{P}_\alpha$, independent across vertices.*

Label vertices using integer labels according to the order in which they enter BP_α so that the root is labelled as 1, the next vertex to be born labeled by 2 and so on. For fixed $t \geqslant 0$, we will view $\mathsf{BP}_\alpha(t)$ as a (random) labelled tree representing the genealogical relationships between all individuals in the population present at time t. See Figures 3.8 and 3.9. Write $|\mathsf{BP}_\alpha(t)|$ for the number of individuals in the tree by time t. Fix $m \geqslant 1$ and define the stopping time

$$\tau_m := \inf \{t : |\mathsf{BP}_\alpha(t)| = m\} . \tag{3.31}$$

Since there are no deaths and each individual reproduces at rate at least $1 + \alpha$, the stopping times $\tau_m < \infty$ a.s. for all $m \geqslant 1$. Now consider the original preferential attachment model where there is no change point. Using properties of the exponential distribution, the following lemma is easy to check and is just a special case of the famous Athreya-Karlin embedding [6].

Lemma 3.4.2. *Viewed as random rooted trees on vertex set* $[n]$ *one has* $\mathsf{BP}_\alpha(\tau_n) \stackrel{d}{=} \mathcal{T}_n$. *In fact the two processes of growing random trees have the same distribution namely*

$$\{\mathsf{BP}_\alpha(\tau_n) : n \geqslant 1\} \stackrel{d}{=} \{\mathcal{T}_n : n \geqslant 1\}\,.$$

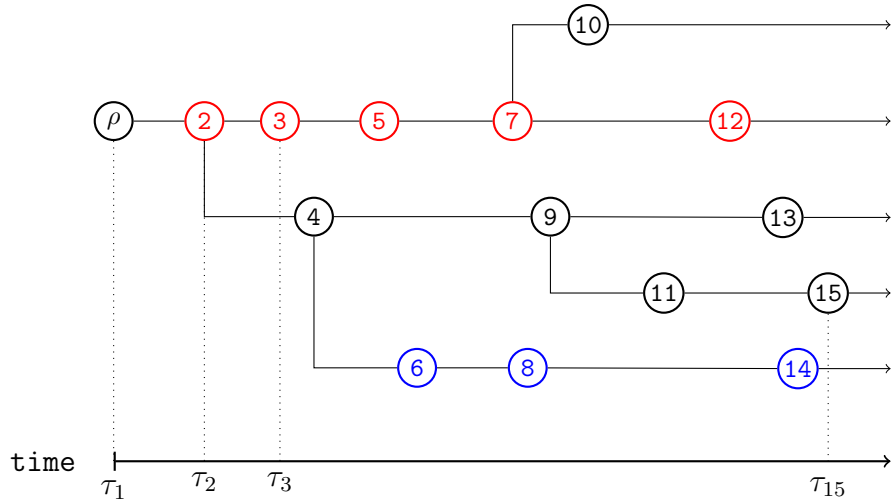


Figure 3.8: The process $\mathsf{BP}_\alpha(\cdot)$ in continuous time starting from the root ρ and stopped at τ_{15}.

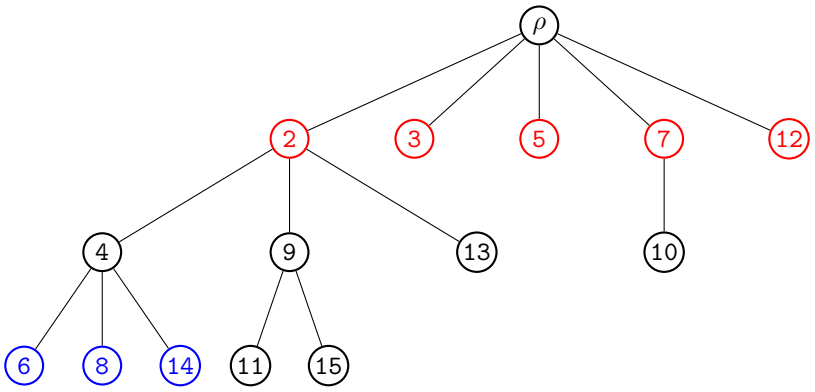


Figure 3.9: The corresponding discrete tree containing only the genealogical information of vertices in $\mathsf{BP}_\alpha(\tau_{15})$.

To construct the variant $\mathcal{T}_n$ where one has a change point, we run $\mathsf{BP}_\alpha(\cdot)$ until time $\tau_{\gamma n}$ (when the original process reaches size γn) and then every vertex changes the way it reproduces. More precisely, after this stopping time, an individual with k children would have reproduced at rate $k + 1 + \alpha$ in the original model but in the change point model

this vertex reproduces at rate $k+1+\beta$ and uses the parameter β instead of α for each subsequent offspring times. Each new vertex v produced after time $\tau_{\gamma n}$ reproduces according to an independent copy of the point process $\mathcal{P}_\beta$. Call the resulting process $\mathsf{BP}^n_{\boldsymbol{\theta}}(\cdot)$ and run the process until time τ_n when the continuous-time process has n individuals. Analogous to (3.31), define the collection of stopping times $\{\tau_m : 1 \leqslant m \leqslant n\}$ by replacing BP_α with $\mathsf{BP}^n_{\boldsymbol{\theta}}$. The following is a simple extension of the previous lemma.

Lemma 3.4.3. *Recall the family of random trees $\{\mathcal{T}_{\boldsymbol{\theta},m} : 1 \leqslant m \leqslant n\}$ generated using the change point preferential attachment model in Section 3.1.2.1. Then,*

$$\{\mathsf{BP}^n_{\boldsymbol{\theta}}(\tau_m) : 1 \leqslant m \leqslant n\} \stackrel{d}{=} \{\mathcal{T}_{\boldsymbol{\theta},m} : 1 \leqslant m \leqslant n\}.$$

Remark 7. Note that the processes $\{\mathcal{T}_{\boldsymbol{\theta},m} : 1 \leqslant m \leqslant n\}$ when one has a change point are **not** nested in a nice manner as growing trees for **different values** of n. Compare this with the original model (without change point) where we can view the entire sequence $\{\mathcal{T}_n : n \geqslant 1\}$ as an increasing family of random trees. In the above construction it will be convenient to couple the processes across different n by using a **single** common branching process BP_α to generate the tree before the change point $\tau_{\gamma n}$ and then let the process evolve independently after the change point for different n using the prescribed dynamics modulated by the attachment parameter β. Further it will be convenient to allow the process $\mathsf{BP}^n_{\boldsymbol{\theta}}$ to continue to grow after time τ_n as opposed to stopping it exactly at time τ_n.

For future reference, for each vertex v, we will use T_v for the time of birth of this vertex into the system. For fixed time t and a vertex v born before time t (namely $T_v \leqslant t$), we write $d_v(t)$ for the number of children of this vertex by time t. Note that for all $v \neq \rho \in \mathsf{BP}^n_{\boldsymbol{\theta}}(t)$, the full degree of v by time t is $d_v(t)+1$.

We will need some simple stochastic calculus calculations below to derive martingales related to processes of interest. Given a process $\{Z(t) : t \geqslant 0\}$ adapted to a filtration $\{\mathcal{F}(t) : t \geqslant 0\}$, we write $\mathbb{E}(dZ(t)|\mathcal{F}(t)) = a(t)dt$ for an adapted process $a(\cdot)$ if $Z(t) - \int_0^t a(s)ds$

is a (local) martingale. Similarly write $\mathrm{Var}(dZ(t)|\mathcal{F}(t)) = b(t)dt$ if the process

$$V(t) := \left(Z(t) - \int_0^t a(s)ds\right)^2 - \int_0^t b(s)ds, \qquad t \geqslant 0,$$

is a local martingale.

Now recall that $\mathsf{BP}_\alpha(\tau_{\gamma n})$ is the random tree before the change point. These random trees are distributed as the original preferential attachment model without change point using attachment dynamics with parameter α. Using (3.1) and recalling that $N_n(k, \gamma n)$ denotes the number of vertices with degree k results in the following.

Lemma 3.4.4. *For each fixed $k \geqslant 1$ we have $N_n(k, \gamma n)/\gamma n \xrightarrow{\text{a.s.}} p_\alpha(k)$, as $n \to \infty$ where $p_\alpha(\cdot)$ is the probability mass function in* (3.1).

Recall that the branching process BP_α is driven by the offspring point process $\mathcal{P}_\alpha$ and $\mathcal{P}_\alpha(t) := \mathcal{P}_\alpha[0, t]$ is the number of points in $[0, t]$. Define the process

$$M_\alpha(t) := e^{-t}\mathcal{P}_\alpha(t) - (1 + \alpha)(1 - e^{-t}), \qquad t \geqslant 0 \tag{3.32}$$

Lemma 3.4.5. *The process $\{M_\alpha(t) : t \geqslant 0\}$ is a martingale with respect to the natural filtration of $\mathcal{P}_\alpha$. In particular*

$$\mathbb{E}(\mathcal{P}_\alpha(t)) = (1 + \alpha)(e^t - 1) \tag{3.33}$$

Proof: Write $\{\mathcal{F}(t) : t \geqslant 0\}$ for the natural filtration of the process. It is enough to show for all $t \geqslant 0$, $\mathbb{E}(dM_\alpha(t)|\mathcal{F}(t)) = 0$. By construction

$$\mathbb{E}(d\mathcal{P}_\alpha(t)|\mathcal{F}(t)) = (1 + \alpha + \mathcal{P}_\alpha(t))dt.$$

Further

$$\mathbb{E}(dM_\alpha(t)|\mathcal{F}(t)) = e^{-t}\,\mathbb{E}(d\mathcal{P}_\alpha(t)|\mathcal{F}(t)) - e^{-t}\mathcal{P}_\alpha(t)dt + (1 + \alpha)e^{-t}dt.$$

Elementary algebra completes the proof. The final assertion regarding (3.33) follows using the martingale property of M_α and the initial condition $\mathcal{P}_\alpha(0) = 0$. ■

The starting point in the analysis of continuous-time branching processes is the so-called Malthusian rate of growth parameter $\lambda > 0$ which solves the equation

$$\int_0^\infty e^{-\lambda t} d\,\mathbb{E}(\mathcal{P}_\alpha(t)) = \int_0^\infty \lambda e^{-\lambda t}\,\mathbb{E}(\mathcal{P}_\alpha(t))dt = 1 \tag{3.34}$$

Where the first equality follows from writing

$$\int_0^\infty e^{-\lambda t} d\,\mathbb{E}(\mathcal{P}_\alpha(t)) = \int_0^\infty \int_0^\infty \lambda e^{-\lambda x} \mathbb{1}\{x \geqslant t\}\, dx\, d\,\mathbb{E}(\mathcal{P}_\alpha(t))$$

and applying Fubini. Using Lemma 3.4.5 now implies

$$\lambda = 2 + \alpha. \tag{3.35}$$

Let T_λ be an exponential random variable with parameter λ independent of $\mathcal{P}_\alpha$ and consider the integer valued random variable $\mathcal{P}_\alpha(T_\lambda)$. Note that (3.34) is equivalent to $\mathbb{E}(\mathcal{P}_\alpha(T_\lambda)) = 1$. Recall that D_α is a random variable with the (non-change point) degree distribution (3.1). It is easy to check that $D_\alpha - 1 \stackrel{d}{=} \mathcal{P}_\alpha(T_\lambda)$. In particular for $\alpha \geqslant 0$,

$$\mathbb{E}(\mathcal{P}_\alpha(T_\lambda) \log^+ \mathcal{P}_\alpha(T_\lambda)) < \infty.$$

Using standard Jagers-Nerman stable age-distribution theory for branching processes [66, 67] now implies the following.

Proposition 3.4.6. *There exists an integrable a.s. positive random variable W_α such that*

$$e^{-(2+\alpha)t} |\mathsf{BP}_\alpha(t)| \stackrel{a.e.,\mathbb{L}^1}{\longrightarrow} W_\alpha.$$

In particular

$$\tau_{\gamma n} - \frac{1}{2+\alpha}\log n \xrightarrow{\text{a.s.}} W'_\alpha, \tag{3.36}$$

for a finite random variable W'_α.

We conclude this section with asymptotics for the amount of " continuous time" where the attachment dynamics using β is valid, namely $\tau_n - \tau_{\gamma n}$. Recall the constant a from (3.5). We will also write $\{\mathcal{F}_n(t) : t \geqslant 0\}$ for the natural filtration of the process $\{\mathsf{BP}^n_{\boldsymbol{\theta}}(t) : t \geqslant 0\}$.

Lemma 3.4.7. *Let* $\Upsilon_n = \tau_n - \tau_{\gamma n}$ *denote the time after the change point in the continuous-time embedding. Then*

$$\sqrt{n}(\Upsilon_n - a) \xrightarrow{\text{w}} \frac{1}{2+\beta}\sqrt{\frac{1-\gamma}{\gamma}}Z,$$

as $n \to \infty$. *Here* Z *is a standard normal random variable.*

Proof: Note that $\mathsf{BP}^n_{\boldsymbol{\theta}}(\cdot)$ is a Markov process. Further for $t \geqslant \tau_{\gamma n}$ conditional on $\mathsf{BP}^n_{\boldsymbol{\theta}}(t)$, the rate at which a new individual is born into the system is given by

$$\begin{aligned}\lambda(t) &:= \sum_{v \in \mathsf{BP}^n_{\boldsymbol{\theta}}(t)} (d_v(t) + 1 + \beta) \\ &= (2+\beta)|\mathsf{BP}^n_{\boldsymbol{\theta}}(t)| - 1,\end{aligned} \tag{3.37}$$

In particular

$$\Upsilon_n \stackrel{d}{=} \sum_{j=\lfloor \gamma n \rfloor}^{n-1} \frac{E_i}{(2+\beta)j - 1}, \tag{3.38}$$

where $\{E_i : i \geqslant 1\}$ is a sequence of *iid* rate one exponential random variables. To finish the proof it will be enough to derive the limiting variance and check Lyapunov's condition. To ease notation we show the equivalent but slightly cleaner formulation:

$$\sqrt{n}\sum_{j=\lfloor \gamma n \rfloor}^{n}\left(\frac{E_j}{j} - \frac{1}{j}\right) \xrightarrow{\text{w}} \sqrt{\frac{1-\gamma}{\gamma}}Z$$

We need the following limit results:

$$\lim_{n\to\infty} \sum_{j=\lfloor \gamma n \rfloor}^{n} \frac{n}{j^2} = \frac{1-\gamma}{\gamma} \quad \text{and} \quad \lim_{n\to\infty} \sum_{j=n\gamma}^{n} \frac{n^2}{j^4} = 0$$

The first limit can be shown by making the change $x = j/n$ in the index of the summation, whereby we obtain

$$\sum_{j=\lfloor n\gamma \rfloor}^{n} \frac{n}{j^2} \quad \to \quad \int_{\gamma}^{1} \frac{1}{x^2} dx = \frac{1-\gamma}{\gamma}$$

To show the second limit write the original sequence as a product of two sequences and show the convergence of the latter (to a finite limit) by following the same recipe as above:

$$\sum_{j=\lfloor n\gamma \rfloor}^{n} \frac{n^2}{j^4} = \frac{1}{n} \cdot \sum_{j=\lfloor n\gamma \rfloor}^{n} \frac{n^3}{j^4} \to 0$$

Then the limiting variance is calculated simply as:

$$\mathbb{E}\, S_n^2 = n \sum_{j=\lfloor n\gamma \rfloor}^{n} \mathbb{E} \left(\frac{E_j}{j} - \frac{1}{j} \right)^2 = \sum_{j=\lfloor n\gamma \rfloor}^{n} \frac{n}{j^2} \;\to\; \frac{1-\gamma}{\gamma}$$

And Lyapunov's condition is similarly shown:

$$\frac{\sum_{j=\lfloor n\gamma \rfloor}^{n} n^2 \; \mathbb{E} \left(\frac{E_j}{j} - \frac{1}{j} \right)^4}{\left(\mathbb{E}\, S_n^2 \right)^2} = \frac{\sum_{j=\lfloor n\gamma \rfloor}^{n} n^2 \; \mathbb{E} \left(\frac{E_j}{j} - \frac{1}{j} \right)^4}{\left(\sum_{j=\lfloor n\gamma \rfloor}^{n} \frac{n}{j^2} \right)^2} = \frac{9 \sum_{j=\lfloor n\gamma \rfloor}^{n} \frac{n^2}{j^4}}{\left(\sum_{j=\lfloor n\gamma \rfloor}^{n} \frac{n}{j^2} \right)^2} \;\to\; 0$$

■

Corollary 3.4.8. *Define* $M_n = \sum_{j=\lfloor n\gamma \rfloor}^{n} E_j/j$. *Then* $M_n = \log \gamma^{-1} + O_p(n^{-1/2})$.

From the above proposition we know that $\sqrt{n} \left(M_n - \sum_{j=\lfloor n\gamma \rfloor}^{n} \frac{1}{j} \right)$ converges weakly to a mean-zero normally distributed random variable X. It is known that

$$\sum_{j=1}^{n} \frac{1}{j} = \log n + c + \epsilon_n$$

where c is the Euler-Mascheroni constant and $\epsilon_n \sim 1/2n$. Therefore $\sum_{j=\lfloor n\gamma \rfloor}^{n} = \log \gamma^{-1} + a_n$ where $a_n \sim C \cdot n^{-1}$ and we may write

$$\sqrt{n}(M_n - \log \gamma^{-1}) \stackrel{\mathrm{d}}{\Longrightarrow} X$$

The weak limit X is tight so $\sqrt{n}(M_n - \log \gamma^{-1})$ is stochastically bounded. ■

Using the distributional characterization in (3.38), one can show the following tail bound on Υ_n. The proof is standard, using a Chernoff bound with the fact that the random variable on the left-hand side is subexponential.

Lemma 3.4.9. *For any $\kappa > 0$ there exists $N = N(\kappa) < \infty$ such that for all $n > N(\kappa)$,*

$$\mathbb{P}\left(|\Upsilon_n - a| > \frac{1}{n^{1/3}}\right) \leqslant \frac{1}{n^{\kappa}}.$$

In particular by Borel-Cantelli, $\mathbb{P}\left(|\Upsilon_n - a| \leqslant n^{-1/3} \text{ eventually }\right) = 1$.

Here the bound $n^{-1/3}$ was arbitrary. An upper bound of $n^{-(1/2-\delta)}$ with any $\delta > 0$ would result in identical result as above. We fix $n^{-1/3}$ for definiteness.

We end this section by defining the Yule process. Properties of this process will be needed in the next few sections.

Definition 3.4.10 (Rate ν Yule process)**.** *Fix $\nu > 0$. A rate ν Yule process is a pure birth process $\{Y_\nu(t) : t \geqslant 0\}$ with $Y_\nu(0) = 1$ and where the rate of birth of new individuals is proportional to size of the current population. More precisely*

$$\mathbb{P}(Y_\nu(t+) - Y_\nu(t)|\mathcal{F}(t)) := \nu Y_\nu(t)dt + o(dt),$$

where $\{\mathcal{F}(t) : t \geqslant 0\}$ is the natural filtration of the process.

The following is a standard property of the Yule process, see e.g. [84, Section 2.5].

Lemma 3.4.11. *Fix time $t > 0$ and rate $\nu > 0$. Then the random variable $Y_\nu(t)$, namely the number of individuals in the population by time t has a Geometric distribution with parameter $p = e^{-\nu t}$ namely*

$$\mathbb{P}(Y_\nu(t) = k) = e^{-\nu t}(1 - e^{-\nu t})^{k-1}, \qquad k \geqslant 1.$$

Proof: Define $T_n =$ the time at which the population size of the rate-ν Yule process jumps from n to $n + 1$. Then clearly

$$\mathbb{P}(Y_\nu(t) > n) = \mathbb{P}(T_n \leqslant t)$$

By definition we have $T_n \stackrel{\mathrm{d}}{=} W_1 + \ldots + W_n$ where the $W_i \sim \exp(i\nu)$ are independent. But in fact, we also have that

$$T_n \stackrel{\mathrm{d}}{=} \max_{1 \leqslant i \leqslant n} Y_i$$

where the Y_i are independent $\exp(\nu)$ random variables. The proof is complete if we can show this, as $\mathbb{P}(\max_{1 \leqslant i \leqslant n} Y_i \leqslant t) = (1 - e^{-t\alpha})^n$ which is the CCDF of a geometric$(e^{-t\alpha})$ random variable. To see why the above is true, write $\{Y_{(i)} : 1 \leqslant i \leqslant n\}$ for the order statistics of $\{Y_i : 1 \leqslant i \leqslant n\}$, with $Y_{(1)} = \min_{1 \leqslant i \leqslant n} Y_i$ and $Y_{(n)} = T_n$.

Clearly $Y_{(1)} \sim \exp(n\nu)$. Now by memorylessness, the additional waiting times for the $n - 1$ *remaining* Y_i's are still $\exp(\nu)$ and independent. Therefore $Y_{(2)} \sim \exp((n - 1)\nu)$. Iterating n times completes the proof. ■

3.4 3.4.2. Convergence of the degree distribution

In this section we will prove Theorem 3.2.1.

3.4 3.4.2.1. Overview of the proof

Recall the random variables D_α and $D_{\boldsymbol{\theta}}$ set out in Section 3.1.3. D_α is a random variable with the same distribution as that of the limiting degree distribution of the graph without change point, and $D_{\boldsymbol{\theta}}$ is a random variable with the same distribution as that of the limiting

degree distribution of the with-changepoint graph driven by parameter set $\boldsymbol{\theta}$. It will be easier to deal with the random variable $D^{\mathsf{out}}_{\boldsymbol{\theta}} := D_{\boldsymbol{\theta}} - 1$.

In this proof we will lean on the fact that the limiting degree distribution for the with-changepoint model can be seen as splitting into two parts:

1. For a vertex born into the system before time γ, its family line evolves according to parameter α for some portion of its life and then a mix of the influence of parameters α and β once we pass the change point.

2. For a vertex born into the system after time γ, its family line evolves according to parameter β only.

The probability that a vertex is born into the system before the change point is γ. The probability that a vertex is born into the system after the change point is $1-\gamma$. Now letting BC stand for "before changepoint" and AC stand for "after changepoint" everywhere they appear, the distribution of $D^{\mathsf{out}}_{\boldsymbol{\theta}}$ can be written succinctly as:

(a) with probability γ, $D^{\mathsf{out}}_{\boldsymbol{\theta}} := Y_{\mathsf{BC}}$ where $Y_{\mathsf{BC}} = D_\alpha - 1 + N_\beta^{D_\alpha}[0, a]$;

(b) with probability $1-\gamma$, $D^{\mathsf{out}}_{\boldsymbol{\theta}} = Y_{\mathsf{AC}}$ where $Y_{\mathsf{AC}} := N_\beta[0, \mathsf{Age}]$.

For a vertex born before the changepoint, the decomposition of Y_{BC} is explained like this. D_α represents the cumulative effect of the evolution due to α up until the change point time $\tau_{n\gamma}$. Now recall that the constant a represents the amount of *continuous time* it takes for the network to get from its size at the change point $n\gamma$ to its final size n. Then $N_\beta^{D_\alpha}[0, a]$ represents the cumulative effect of the evolution due to β for the rest of the life of the network. Note that this quantity is $N_\beta^{D_\alpha}$, not N_β, so it takes into account the initial amount of influence left by its time in the pre-changepoint regime. Therefore, the total effect on the degree distribution due to vertices born before γ is $Y_{\mathsf{BC}} := D_\alpha - 1 + N_\beta^{D_\alpha}[0, a]$.

For a vertex born after the change point, it evolves solely according to β for the total amount of time it has been alive by the end of the graph process, which is equal to its age, hence the notation Age. Therefore $Y_{\mathsf{AC}} := N_\beta[0, \mathsf{Age}]$.

To set up the convergence result we will need to define some empirical degree counts which will converge to the above limits. Recall that for any time t and vertex v born before time t, $d_v(t)$ denotes the number of children (out-degree) of vertex v at time t. For fixed $k \geqslant 0$ define

$$\bar{N}_n^{\mathsf{BC}}(k) := \sum_{v \in \mathsf{BP}_{\boldsymbol{\theta}}(\tau_n)} \mathbb{1}\left\{T_v \leqslant \tau_{\gamma n}, d_v(\tau_n) \geqslant k\right\}, \tag{3.39}$$

and

$$\bar{N}_n^{\mathsf{AC}}(k) := \sum_{v \in \mathsf{BP}_{\boldsymbol{\theta}}(\tau_n)} \mathbb{1}\left\{T_v > \tau_{\gamma n}, d_v(\tau_n) \geqslant k\right\}. \tag{3.40}$$

In words, $\bar{N}_n^{\mathsf{BC}}(k)$ is the number of vertices that were born before the change point and have out-degree at least k by time τ_n (thus in the tree $\mathcal{T}_{\boldsymbol{\theta},n}$) whilst $\bar{N}_n^{\mathsf{AC}}(k)$ is defined similarly but for vertices born after the change point $\tau_{\gamma n}$. The following proposition is equivalent to Theorem 3.2.1.

Proposition 3.4.12. *Fix $k \geqslant 0$. Then we have*

$$\frac{\bar{N}_n^{\mathsf{BC}}(k)}{n} \xrightarrow{\mathrm{P}} \gamma\, \mathbb{P}(Y_{\mathsf{BC}} \geqslant k), \qquad \frac{\bar{N}_n^{\mathsf{AC}}(k)}{n} \xrightarrow{\mathrm{P}} (1-\gamma)\, \mathbb{P}(Y_{\mathsf{AC}} \geqslant k), \tag{3.41}$$

as $n \to \infty$.

The rest of this section deals with proving this proposition.

3.4 3.4.2.2. Analysis of $\bar{N}_n^{\mathsf{BC}}(\cdot)$:

We start with the easier case. We will need some more notation. For fixed $j, k \geqslant 0$, define $\bar{N}_n^{\mathsf{BC}}(j : k)$ for the number of vertices that were born before the change point $\tau_{\gamma n}$ with out-degree exactly j at time $\tau_{\gamma n}$ that end up with at least k children by time τ_n. Note that

$$\sum_{j \geqslant k} \bar{N}_n^{\mathsf{BC}}(j : k) = N_n(k+1, \gamma n) = \bar{N}_n^{\mathsf{BC}}(k)$$

namely the number of vertices with total degree $k+1$ (thus out-degree k) in the tree before change point $\mathcal{T}_{\gamma n}$. Recall that the goal of this section is to prove the convergence

$$\bar{N}_n^{\mathsf{BC}}(k)/n \xrightarrow{\text{P}} \gamma\, \mathbb{P}(Y_{\mathsf{BC}} \geqslant k) \tag{3.42}$$

This is a statement involving all vertices born before the change point. However, because a vertex born before the change point is subject to two different growth regimes (pre- and post-change), it will be easier to condition on the state of the vertex at the exact moment of the change so we can break the analysis of 3.42 into two parts—the effect due to the BC regime and the effect due to the AC regime. To do this we will condition on the event that a vertex *has out-degree exactly equal to j at time $\tau_{n\gamma}$*. Proving the convergence on the conditioning sets will then imply the main result.

Conditional on this event, the left-hand side becomes $\bar{N}_n^{\mathsf{BC}}(j:k)/n$. The right-hand side a bit more complicated. The limit $\mathbb{P}(Y_{\mathsf{BC}} \geqslant k)$ represents the limiting probability of the event that a BC vertex has degree at least k by time τ_n. The conditioned version of this event is the probability that a BC vertex *with out-degree $= j$ at time $\tau_{n\gamma}$* will grow to have degree at least k by time τ_n. This probability is the product of

1. $\mathbb{P}(D_\alpha^{\mathsf{out}} = j)$, the probability that a BC vertex will have out-degree exactly j at time $\tau_{n\gamma}$.

2. $\mathbb{P}(\mathcal{P}_\beta^{j+1}[0,a] \geqslant k-j)$, the probability that said vertex will acquire at least $k-j$ degrees in the a duration of time between the time of the change point $\tau_{n\gamma}$ and the end of the process τ_n, during which it is subject to growth according to parameter β.

Therefore it is enough to show for each fixed $0 \leqslant j \leqslant k$,

$$\frac{\bar{N}_n^{\mathsf{BC}}(j:k)}{n} \xrightarrow{\text{a.s.}} \gamma\, \mathbb{P}(D_\alpha^{\mathsf{out}} = j)\, \mathbb{P}(\mathcal{P}_\beta^{j+1}[0,a] \geqslant k-j). \tag{3.43}$$

We start with the following simple lemma.

Lemma 3.4.13. *Fix $0 < p, q < 1$, a sequence of non-negative integer valued random variables $\{N_n : n \geqslant 1\}$ and a sequence $\{q_n : n \geqslant 1\} \in [0,1]$. Conditional on N_n, let S_n be a Binomial(N_n, q_n) random variable. Further suppose that*

$$\frac{N_n}{n} \xrightarrow{\text{a.s.}} p, \qquad q_n \to q.$$

Then $S_n/n \xrightarrow{\text{a.s.}} pq$.

Proof: We assume we work on a rich enough probability space where we can couple $\{S_n : n \geqslant 1\}$ with a sequence $\left\{\tilde{S}_n : n \geqslant 1\right\}$ where $\tilde{S}_n$ is Binomial(np, q_n) such that $|S_n - \tilde{S}_n| \leqslant |N_n - np|$. Standard exponential tail bounds for the Binomial distribution coupled with Borel-Cantelli and the hypothesis of the lemma imply that $\tilde{S}_n/n \xrightarrow{\text{a.s.}} pq$. Since $|S_n - \tilde{S}_n|/n \leqslant |N_n/n - p|$, again using the hypothesis of the lemma completes the proof. ■

We proceed with the proof. Recall the definition of the random variable $\bar{N}_n^{\mathsf{BC}}(j : k)$ at the beginning of this section. In the same vein, for each $s \geqslant 0$ define $\bar{Z}_n^{\mathsf{BC}}((j : k), s)$ for the number of vertices born before the change point $\tau_{\gamma n}$ such that at $\tau_{\gamma n}$ they have out-degree exactly j and further by time $\tau_{\gamma n} + s$ they have degree at least k. Then note that conditional on the information at time $\tau_{\gamma n}$,

$$\bar{Z}_n^{\mathsf{BC}}((j : k), s) \stackrel{d}{=} \text{Bin}(N_n(j+1, \gamma n), \mathbb{P}(\mathcal{P}_\beta^{j+1}[0, s] \geqslant k - j)) \tag{3.44}$$

Further the random variables of interest $\bar{N}_n^{\mathsf{BC}}(j : k) = \bar{Z}_n^{\mathsf{BC}}((j : k), \Upsilon_n)$ where Υ_n is as in Lemma 3.4.7. Thus writing $a_n^+ = a + n^{-1/3}$ and $a_n^- = a - n^{-1/3}$ and using Lemma 3.4.9,

$$\bar{Z}_n^{\mathsf{BC}}((j : k), a_n^-) \leqslant \bar{N}_n^{\mathsf{BC}}(j : k) \leqslant \bar{Z}_n^{\mathsf{BC}}((j : k), a_n^+) \text{ eventually a.s.} \tag{3.45}$$

Using the Binomial convergence lemma 3.4.13 and noting that by Lemma 3.4.4 and choice of a_n^+, a_n^-, the hypothesis of this Lemma are satisfied, implies that

$$\frac{\bar{Z}_n^{\mathsf{BC}}((j:k), a_n)}{n} \xrightarrow{\text{a.s.}} \gamma\, \mathbb{P}(D_\alpha^{\mathsf{out}} = j)\, \mathbb{P}(\mathcal{P}_\beta^{j+1}[0, a] \geqslant k - j),$$

where take a_n as either a_n^+ or a_n^-. Now using (3.45) proves (3.43). This completes the analysis of $\bar{N}_n^{\mathsf{BC}}(\cdot)$.

■

3.4 3.4.2.3. Analysis of $\bar{N}_n^{\mathsf{AC}}(\cdot)$:

Similar to the technique in the previous section, we will begin by simplifying the convergence we want to show using conditioning. However we still need some notation first. Fix $k \geqslant 0$ and define the function

$$g_k(u) := \mathbb{P}(\mathcal{P}_\beta[0, u] \geqslant k), \qquad u \geqslant 0. \tag{3.46}$$

Here $\mathcal{P}_\beta$ is the offspring point process with attachment parameter β. The convergence we want to show is

$$\bar{N}_n^{\mathsf{AC}}(k)/n \xrightarrow{\text{P}} (1 - \gamma)\, \mathbb{P}(Y_{\mathsf{AC}} \geqslant k) \tag{3.47}$$

Since $Y_{\mathsf{AC}} = \mathcal{P}_\beta[0, \mathsf{Age}]$, we can make the picture clearer by conditioning on the possible ages of each vertex in the AC regime. Actually, it turns out to be easier and equivalent to condition on the *time of birth* of each vertex. Write $\{\text{Born} = u\}$ for the event that a vertex in the AC regime is born exactly u time units after the change point. Then the right-hand side limiting probability in 3.47 can be written

$$\int_0^a (1 - \gamma)\, \mathbb{P}(N_\beta[0, a - u] \geqslant k \,|\, \text{Born} = u)\ \mathbb{P}(\text{Born} = u)\, du \tag{3.48}$$

The probability that an AC vertex is born exactly u time units into the AC regime is derived using Proposition 3.4.6. That result implies that the number of individuals in the BC regime after u time units has passed is proportional to $e^{(2+\beta)u}$. Therefore since the total amount of time available in the AC regime is a,

$$\mathbb{P}(\text{Born} = u) = \frac{e^{(2+\beta)u}}{\int_0^a e^{(2+\beta)u} du} = \frac{(2+\beta)e^{(2+\beta)u}}{e^{(2+\beta)a} - 1} \tag{3.49}$$

Then using the definition of a from (3.5) and simplifying, we see that to prove the second assertion of (3.41), it is equivalent to show

$$\frac{\bar{N}_n^{\mathsf{AC}}(k)}{n} \xrightarrow{\text{P}} \gamma \int_0^a (2+\beta)e^{(2+\beta)u} g_k(a-u)du. \tag{3.50}$$

Let us begin. For $s \geqslant 0$, define $\bar{Z}_n^{\mathsf{AC}}(k,s)$ for the number of individuals born in the interval $[\tau_{\gamma n}, \tau_{\gamma n} + s]$ such that by time $\tau_{\gamma n} + s$, these vertices have at least k children. Then note that $\bar{N}_n^{\mathsf{AC}}(k) = \bar{Z}_n^{\mathsf{AC}}(k, \Upsilon_n)$. Mimicking the proof of $N_n^{\mathsf{BC}}(k)$, it is enough to show that

$$\frac{\bar{Z}_n^{\mathsf{AC}}(k,a_n)}{n} \xrightarrow{\text{P}} \gamma(2+\beta) \int_0^a e^{(2+\beta)u} g_k(a-u)du, \tag{3.51}$$

where a_n is either the sequence $a_n^- = a - n^{-1/3}$ or $a_n^+ = a + n^{-1/3}$. To ease notation we will just work with the sequence $a_n = a$. The entire proof goes through by replacing a in the steps below by a_n.

We start with a few preliminary results. The first result describes strong concentration results of the growth of the number of individuals in $\mathsf{BP}_{\boldsymbol{\theta}}^n$ in the interval $[\tau_{\gamma n}, \tau_{\gamma n} + s]$. Define the process

$$\mathscr{Z}_n(u) := |\mathsf{BP}_{\boldsymbol{\theta}}^n(\tau_{\gamma n} + u)|, \qquad 0 \leqslant u \leqslant a. \tag{3.52}$$

Proposition 3.4.14. *There exists a constant $C < \infty$ such that for all n,*

$$\mathbb{P}\left(\sup_{0 \leqslant u \leqslant a} |\mathscr{Z}_n(u) - n\gamma e^{(2+\beta)u}| > \sqrt{n \log n} \right) \leqslant \frac{C}{\log n}.$$

Proof: The plan is to use Doob's L^2-maximal inequality for continuous-time martingales (see e.g. [74, Chapter 1.9]). For this we will need to derive martingales related to the process $\mathscr{Z}_n(\cdot)$. Throughout we will write $\{\mathcal{F}_t^n : 0 \leqslant t \leqslant a\}$ for the filtration $\{\mathsf{BP}_{\boldsymbol{\theta}}(\tau_{\gamma n} + t) : 0 \leqslant t \leqslant a\}$. Recall from the rate description in (3.37) that $\mathscr{Z}_n(\cdot)$ is a pure birth process such for any $t \geqslant 0$, conditional on $\mathcal{F}_t^n$, $\mathscr{Z}_n(t) \rightsquigarrow \mathscr{Z}_n(t)+1$ at rate $(2+\beta)\mathscr{Z}_n(t)-1$. Arguing as in the proof of Lemma 3.4.5 it is easy to check that the process

$$M_1(t) := \left(e^{-(2+\beta)t}\,\mathscr{Z}_n(t) - n\gamma\right) - \frac{e^{-(2+\beta)t}-1}{2+\beta}, \qquad 0 \leqslant t \leqslant a, \tag{3.53}$$

is a mean-zero martingale. This in particular gives that

$$e^{-(2+\beta)t}\,\mathbb{E}(\mathscr{Z}_n(t)) = n\gamma + \frac{e^{-(2+\beta)t}-1}{2+\beta}, \qquad 0 \leqslant t \leqslant a. \tag{3.54}$$

By Doob's L^2-maximal inequality applied to the process $M_1(\cdot)$ we have for any $\lambda > 0$,

$$\mathbb{P}\left(\sup_{0\leqslant t\leqslant a}\left|\left(e^{-(2+\beta)t}\,\mathscr{Z}_n(t) - n\gamma\right) - \frac{e^{-(2+\beta)t}-1}{2+\beta}\right| \geqslant \lambda\right) \leqslant \frac{\mathbb{E}(M_1^2(a))}{\lambda^2}. \tag{3.55}$$

If we can show there exists a constant $C < \infty$ such that $\mathbb{E}(M_1^2(a)) \leqslant Cn$, using $\lambda = .5\sqrt{n\log n}$ and algebraic manipulation of (3.55) completes the proof. So let us now derive this bound on $\mathbb{E}(M_1^2(a))$.

First squaring the expression in (3.53), expanding and using (3.54) gives for $t \geqslant 0$,

$$\mathbb{E}(M_1^2(t)) = \mathbb{E}\left(e^{-(2+\beta)t}\,\mathscr{Z}_n(t) - n\gamma\right)^2 - \left(\frac{e^{-(2+\beta)t}-1}{2+\beta}\right)^2. \tag{3.56}$$

Thus we need to understand the evolution of the process $\mathscr{Z}_n^2(\cdot)$. Again using the rate description of $\mathscr{Z}_n$, this process undergoes a change

$$\Delta\mathscr{Z}_n^2(t) := \mathscr{Z}_n^2(t+) - \mathscr{Z}_n^2(t) = (1 + 2\mathscr{Z}_n(t)),$$

at rate $(2+\beta)\mathscr{Z}_n(t) - 1$. Using this one may check that the following process on $[0, a]$

$$M_2(t) := e^{-2(2+\beta)t}\,\mathscr{Z}_n^2(t) - \int_0^t e^{-2(2+\beta)s}\beta\mathscr{Z}_n(s)ds - \frac{e^{-2(2+\beta)t}}{2(2+\beta)}, \tag{3.57}$$

is also a martingale. In particular since first moments are conserved,

$$\mathbb{E}(e^{-2(2+\beta)}\,\mathscr{Z}_n^2(t)) = n^2\gamma^2 + \int_0^t \beta e^{-2(2+\beta)s}\,\mathbb{E}(\mathscr{Z}_n(s))ds - \frac{e^{-2(2+\beta)t} - 1}{2(2+\beta)}. \tag{3.58}$$

Using (3.54) shows that there exists a constant C such that

$$\left|\mathbb{E}(e^{-2(2+\beta)}\,\mathscr{Z}_n^2(t)) - n^2\gamma^2\right| \leqslant n\gamma. \tag{3.59}$$

Expanding the first bracket in (3.56), using (3.54) and (3.59) shows that $\mathbb{E}(M_1^2(a)) \leqslant Cn$ for some constant C. This completes the proof.

■

Now divide the interval $[\tau_{\gamma n}, \tau_{\gamma n} + a]$ into $an^{1/3}$ intervals of length $[n^{-1/3}]$:

$$\left\{\left[\tau_{\gamma n}, \tau_{\gamma n} + \frac{1}{n^{1/3}}\right], \left[\tau_{\gamma n} + \frac{1}{n^{1/3}}, \tau_{\gamma n} + \frac{2}{n^{1/3}}\right], \ldots, \left[\tau_{\gamma n} + \frac{an^{1/3}-1}{n^{1/3}}, \tau_{\gamma n} + \frac{an^{1/3}}{n^{1/3}}\right]\right\},$$

To ease notation, write the above collection as $\left\{\mathcal{I}_i : 0 \leqslant i \leqslant an^{1/3} - 1\right\}$. Further let $\tau_i^n = \tau_{\gamma n} + i/n^{1/3}$ with $\tau_0^n = \tau_{\gamma n}$ so that $\mathcal{I}_i = [\tau_i^n, \tau_{i+1}^n]$.

Now write Birth_i for the collection of vertices that were born in interval $\mathcal{I}_i$ (i.e. the collection of vertices v with birth times $T_v \in \mathcal{I}_i$) and write

$$\mathscr{Z}_n(\mathcal{I}_i) := |\mathsf{Birth}_i| = \mathscr{Z}_n\left(\tau_{i+1}^n\right) - \mathscr{Z}_n\left(\tau_i^n\right),$$

for the number of individuals born in this interval. Then the following is an easy corollary of Proposition 3.4.14.

Corollary 3.4.15. *We have*

$$\mathbb{P}\left(\bigcap_{i=0}^{an^{1/3}-1}\left\{\left|\mathscr{Z}_n(\mathcal{I}_i)-(2+\beta)\gamma n^{2/3}e^{\frac{(2+\beta)i}{n^{1/3}}}\right|<2\sqrt{n\log n}\right\}\right)\to 1,$$

as $n\to\infty$.

For future reference write $\mathscr{G}_n$ for the event above. Namely,

$$\mathscr{G}_n:=\bigcap_{i=0}^{an^{1/3}-1}\left\{\left|\mathscr{Z}_n(\mathcal{I}_i)-(2+\beta)\gamma n^{2/3}e^{\frac{(2+\beta)i}{n^{1/3}}}\right|<2\sqrt{n\log n}\right\} \tag{3.60}$$

Now for each interval $\mathcal{I}_i$, we will partition the vertices born in this interval into two classes:

(a) The collection of good vertices $\mathcal{G}_i$: This consists of all $v\in\mathsf{Birth}_i$ such that they produce **no** children by the end of the interval i.e. vertices v with $T_v\in[\tau_{\gamma n}+i/n^{1/3},\tau_{\gamma n}+(i+1)/n^{1/3}]$ such that by time $\tau_{\gamma n}+(i+1)/n^{1/3}$, vertex v still has no children. Write $\mathscr{Z}_n^{\mathsf{good}}(\mathcal{I}_i)=|\mathcal{G}_i|$ for the number of good vertices in $\mathcal{I}_i$.

(b) The collection of bad vertices $\mathcal{B}_i:=\mathsf{Birth}_i\setminus\mathcal{G}_i$: This consists of all vertices born in $\mathcal{I}_i$ which produce at least one child by time $\tau_{\gamma n}+i/n^{1/3}$. Write $\mathscr{Z}_n^{\mathsf{bad}}(\mathcal{I}_i)=|\mathcal{B}_i|$ for the number of such bad vertices in $\mathcal{I}_i$. Write

$$\mathscr{Z}_n^{\mathsf{bad}}:=\sum_{i=0}^{an^{1/3}-1}\mathscr{Z}_n^{\mathsf{bad}}(\mathcal{I}_i)$$

for the total number of bad vertices.

The general idea is the following. Note that since the intervals are of time length $n^{-1/3}$ but the average time until first birth for each vertex is constant $1/(1-\beta)$, one expects a large proportion of vertices born in the interval $\mathcal{I}_i$ to be good. At the same time however, the process $|\mathsf{BP}_{\boldsymbol{\theta}}(t)|$ is accelerating exponentially fast. Therefore to be completely sure that

bad vertices are under control, we need to precisely calculate the balance of the two forces and show that the bad vertices eventually are squashed out.

This is the content of the next result. Fix a constant C and define the event $B_i^n = \left\{ \mathscr{Z}_n^{\mathsf{bad}}(\mathcal{I}_i) \geqslant Cn^{1/3} \log n \right\}$. These events depend on C but we suppress this in the notation.

Proposition 3.4.16. *The constant $C < \infty$ can be chosen large enough such that $\mathbb{P}(\cup_{i=1}^{an^{1/3}} B_i^n) \to 0$ as $n \to \infty$. In particular for the total number of bad vertices we have $\mathscr{Z}_n^{\mathsf{bad}} = O_P(n^{2/3} \log n)$.*

Proof: Fix an interval $\mathcal{I}_i$. Note that every bad vertex is one of two types:

(a) A vertex that is a direct child of a vertex born before this time interval. Write $\mathcal{D}_n^{\mathsf{bad}}$ for these *direct* bad vertices and write $\mathscr{D}_n^{\mathsf{bad}}(\mathcal{I}_i) = |\mathcal{D}_n^{\mathsf{bad}}|$ for the number of such vertices. Further write $\mathscr{D}_{n,\star}^{\mathsf{bad}}(\mathcal{I}_i)$ for the total number of descendants of direct bad vertices born in the interval $\mathcal{I}_i$ (including the direct bad vertices).

(b) A vertex that is bad and is a child of a vertex born in $\mathcal{I}_i$. Thus the parent of this vertex is necessarily bad.

Thus in particular we have that $\mathscr{Z}_n^{\mathsf{bad}}(\mathcal{I}_i) \leqslant \mathscr{D}_{n,\star}^{\mathsf{bad}}(\mathcal{I}_i)$. Now note that direct bad vertices in $\mathcal{D}_n^{\mathsf{bad}}$ are created via the following steps:

(i) A descendant (maybe good or bad) of a vertex born before $\mathcal{I}_i$ is born into the system. The number of such individuals $\mathscr{R}_n(\mathcal{I}_i) \leqslant \mathscr{Z}_n(\mathcal{I}_i)$, the total number of individuals born in the interval $\mathcal{I}_i$. Using Corollary 3.4.15, there exists a constant C such that whp as $n \to \infty$, *for all* the intervals $0 \leqslant i \leqslant an^{1/3} - 1$, $\mathscr{R}_n(\mathcal{I}_i) \leqslant Cn^{2/3}$.

(ii) Conditional on **all these** descendants of vertices born before $\mathcal{I}_i$, such a descendant has to give birth to one individual in the interval $[i/n^{1/3}, (i+1)/n^{1/3}]$. Recall that the time to give birth to the first child is an exponential random variable E_1 with rate $(2+\beta)$. Thus the probability of birthing this first child is bounded by

$$p_n = \mathbb{P}(E_1 \leqslant n^{-1/3}) \sim \frac{2+\beta}{n^{1/3}}.$$

Further by construction none of these vertices can have a parent child relationship and thus their offspring lineages evolve independently.

In particular, conditional on all descendants of vertices born before time interval $\mathcal{I}_i$,

$$\mathscr{D}_n^{\mathsf{bad}}(\mathcal{I}_i) \leqslant_{\mathsf{st}} \mathrm{Bin}(\mathscr{R}_n(\mathcal{I}_i), p_n) \tag{3.61}$$

Here st denotes stochastic domination. Thus using Corollary 3.4.15, (3.61) and standard tail bounds for the Binomial distribution implies that there exists a constant $C < \infty$ such that

$$\mathbb{P}(\mathscr{D}_n^{\mathsf{bad}}(\mathcal{I}_i) \leqslant Cn^{1/3} \log n \ \forall 0 \leqslant i \leqslant an^{1/3} - 1) \to 1, \tag{3.62}$$

as $n \to \infty$.

Let us now complete the analysis of $\mathscr{D}_{n,\star}^{\mathsf{bad}}(\mathcal{I}_i)$. Let us start with the evolution of descendants of a single bad *direct* vertex after it gives birth to its child. This process then starts reproducing at rate $2 + \beta + 1 + \beta = 3 + 2\beta$. Further whenever a new vertex is added to the system, the rate of production increases by at most $2 + \beta$. Thus writing $K = \lfloor 3 + 2\beta \rfloor$ and $\nu = 2 + \beta$, the number of descendants of such a bad vertex can be bounded by a rate ν Yule process (see Definition 3.4.10) that starts with K individuals at time zero. Write $\left\{Y_\nu^K(t) : t \geqslant 0\right\}$ for such a process. Thus the number of descendants of such a bad vertex in the time interval $[\tau_{\gamma n} + i/n^{1/3}, \tau_{\gamma n} + (i+1)/n^{1/3}]$ can be stochastically bounded by $Y_\nu^K(n^{-1/3})$. In particular, conditional on $\mathcal{D}_n^{\mathsf{bad}}(\mathcal{I}_i)$,

$$\mathscr{D}_{n,\star}^{\mathsf{bad}}(\mathcal{I}_i) \leqslant_{\mathsf{st}} \sum_{j=1}^{\mathscr{D}_n^{\mathsf{bad}}(\mathcal{I}_i)} Y_\nu^{K,(j)}(n^{-1/3}). \tag{3.63}$$

Here $\left\{Y_\nu^{K,(j)}(\cdot) : j \geqslant 1\right\}$ are an *iid* collection of Yule processes with distribution $Y_\nu^K(\cdot)$. Using the explicit distribution of the Yule process at a fixed time (Lemma 3.4.11), it is easy to

check that given constant $C > 0$ we can find $A > 0$ such that

$$\mathbb{P}\left(\mathscr{D}^{\mathsf{bad}}_{n,\star}(\mathcal{I}_i) \geqslant 10KCn^{1/3}\log n \middle| \mathscr{D}^{\mathsf{bad}}_{n}(\mathcal{I}_i) \leqslant Cn^{1/3}\log n\right) \leqslant \exp(-An^{1/3}). \tag{3.64}$$

Using this exponential bound with (3.62) completes the proof. ■

We now proceed with the proof of (3.51). For $0 \leqslant i \leqslant an^{1/3} - 1$, let $Z^{\mathsf{good}}_n(k, a : \mathcal{I}_i)$ be the number of **good** vertices in Birth_i which have at least k children by time a. Then note that conditional on $\mathsf{BP}^n_{\boldsymbol{\theta}}(\tau^n_{i+1})$,

$$Z^{\mathsf{good}}_n(k, a : \mathcal{I}_i) \stackrel{d}{=} \mathrm{Bin}\left(\mathscr{Z}^{\mathsf{good}}_n(\mathcal{I}_i), g_k\left(a - \frac{i+1}{n^{1/3}}\right)\right). \tag{3.65}$$

Define the events

$$G^n_i := \left\{\left|Z^{\mathsf{good}}_n(k, a : \mathcal{I}_i) - \gamma(2+\beta)n^{2/3}e^{\frac{(2+\beta)i}{n^{1/3}}} g_k\left(a - \frac{i+1}{n^{1/3}}\right)\right| < Cn^{1/3}\log n\right\}$$

Proposition 3.4.17. *There exists a constant $C < \infty$ such that $\mathbb{P}\left(\cap_{i=1}^{an^{1/3}} G^n_i\right) \to 1$ as $n \to \infty$.*

Proof: Note that $\mathscr{Z}^{\mathsf{good}}_n(\mathcal{I}_i) = \mathscr{Z}_n(\mathcal{I}_i) - \mathscr{Z}^{\mathsf{bad}}_n(\mathcal{I}_i)$. Combining Corollary 3.4.15 with Proposition 3.4.16 implies that

$$\mathbb{P}\left(\bigcap_{i=0}^{an^{1/3}-1}\left\{\left|\mathscr{Z}^{\mathsf{good}}_n(\mathcal{I}_i) - (2+\beta)\gamma n^{2/3}e^{\frac{(2+\beta)i}{n^{1/3}}}\right| < 3\sqrt{n\log n}\right\}\right) \to 1,$$

Now using the distributional identity (3.65) and standard tail bounds for the Binomial distribution completes the proof.

■

We are finally in a position to complete the proof of (3.51). First note that

$$\sum_{i=0}^{an^{1/3}-1} Z^{\mathsf{good}}_n(k, a : \mathcal{I}_i) \leqslant \bar{Z}^{\mathsf{AC}}_n(k, a) \leqslant \sum_{i=0}^{an^{1/3}-1} Z^{\mathsf{good}}_n(k, a : \mathcal{I}_i) + \mathscr{Z}^{\mathsf{bad}}_n. \tag{3.66}$$

Using Proposition 3.4.16 $n^{-1}\mathscr{Z}_n^{\mathsf{bad}} \xrightarrow{\mathrm{P}} 0$. Using Proposition 3.4.17

$$\begin{aligned}\frac{\sum_{i=1}^{an^{1/3}} Z_n^{\mathsf{good}}(k, a : \mathcal{I}_i)}{n} &\sim \frac{\gamma(2+\beta)}{n^{1/3}} \sum_{i=0}^{an^{1/3}-1} e^{\frac{(2+\beta)i}{n^{1/3}}} g_k\left(a - \frac{i+1}{n^{1/3}}\right) \\ &\to \gamma(2+\beta)\int_0^a e^{(2+\beta)u} g_k(a-u)du.\end{aligned}$$

This completes the proof of (3.41) and thus the assertion of the convergence of the degree distribution of the model to the asserted limit in Theorem 3.2.1.

■

We conclude this section with a related result regarding the evolution of the degree distribution. This follows by directly modifying the proof above. Recall the definitions of $N_n(k,m)$ and $\hat{N}_n(k,t)$ from Section 3.2.2. For future use define for each $k \geqslant 1$ and $0 \leqslant t \leqslant 1$

$$N_{n,\geqslant}(k,m) = \sum_{j \geqslant k} N_n(j,m), \qquad \hat{N}_{n,\geqslant}(k,t) = \sum_{j \geqslant k} \hat{N}_n(j,t), \tag{3.67}$$

namely the number of vertices with degree at least k respectively at discrete time m and at time t when we rescale time by n. Write $\hat{q}_{\geqslant}^{(n)}(k,t) = \hat{N}_{n,\geqslant}(k,t)/n$. Note that since we divide by n and not nt in this expression we have $\sum_{k=1}^{\infty} \hat{q}_{\geqslant}^{(n)}(k,t) = t$. Now note that by Lemma 3.4.4 we have for each fixed $0 < t \leqslant \gamma$,

$$\hat{p}^{(n)}(k,t) \xrightarrow{\mathrm{P}} p_\alpha(k) = p^{(\infty)}(k,\gamma), \tag{3.68}$$

where $p_\alpha(k)$ as in (3.1) is the limiting degree distribution with no change point. For $\gamma \leqslant t \leqslant 1$, analogous to the definition of a in (3.5) define

$$a(t) := \frac{1}{2+\beta} \log \frac{t}{\gamma} \tag{3.69}$$

Analogous to the definition of $D_{\boldsymbol{\theta}}$ in Section 3.1.3, define $D_{\boldsymbol{\theta}}(t)$ by replacing a by $a(t)$ throughout the construction. Thus $D_{\boldsymbol{\theta}} = D_{\boldsymbol{\theta}}(1)$. Let

$$p^{(\infty)}(k,t) := \mathbb{P}(D_{\boldsymbol{\theta}}(t) = k), \qquad k \geqslant 1, \;\; \gamma \leqslant t \leqslant 1. \tag{3.70}$$

Let $p^{(\infty)}_{\geqslant}(k,t) = \mathbb{P}(D_{\boldsymbol{\theta}}(t) \geqslant k)$. For $0 \leqslant t \leqslant 1$, let $q^{(\infty)}_{\geqslant}(k,t) = tp^{(\infty)}_{\geqslant}(k,t)$.

Proposition 3.4.18. *For all $k \geqslant 1$ we have*

$$\sup_{0 \leqslant t \leqslant 1} |\hat{q}^{(n)}_{\geqslant}(k,t) - q^{(\infty)}_{\geqslant}(k,t)| \xrightarrow{\mathrm{P}} 0,$$

as $n \to \infty$.

Proof: For fixed $t \geqslant \gamma$, define the stopping time

$$\tau_{tn} = \inf \left\{ s : |\mathsf{BP}^n_{\boldsymbol{\theta}}(s)| = tn \right\},$$

namely the first time that the continuous-time embedding reaches size tn. Note that at this time, the corresponding tree has distribution $\mathcal{T}_{tn}$. Write $\Upsilon_n(t) = \tau_{tn} - \tau_{\gamma n}$ for the amount of (continuous) time it takes for the process to reach this size after the change point. Then note that by Proposition 3.4.14 we can choose an appropriate constant $C < \infty$ such that

$$\mathbb{P}\left(\sup_{\gamma \leqslant t \leqslant 1} |\Upsilon(t) - a(t)| \leqslant C\sqrt{\frac{\log n}{n}} \right) \to 1, \tag{3.71}$$

as $n \to \infty$, where $a(t)$ is as defined in (3.69). Repeating the above proof for the convergence of degree distribution and replacing a by $a(t)$ throughout the argument shows that for each $t \geqslant \gamma$ $\hat{N}_{n,\geqslant}(k,t)/nt \xrightarrow{\mathrm{P}} \mathbb{P}(D_{\boldsymbol{\theta}}(t) \geqslant k)$. Combining this with (3.68) implies that we have pointwise convergence $\hat{q}^{(n)}_{\geqslant}(k,t) \to q^{(\infty)}_{\geqslant}(k,t)$. Now note that for each fixed n, the function $\hat{q}^{(n)}_{\geqslant}(k,\cdot)$ is non-decreasing on $[0,1]$ while the limit function is also monotonically increasing and continuous (and thus uniformly continuous). Given $\varepsilon > 0$, fix $\delta > 0$ such that for any

$t, s \in [0,1]$ with $|t-s| < \delta$,

$$|q_{\geqslant}^{(\infty)}(k,t) - q_{\geqslant}^{(\infty)}(k,s)| < \frac{\varepsilon}{4}.$$

Divide $[0,1]$ into intervals $\{[i\delta,(i+1)\delta]\}$ for $1 \leqslant i \leqslant 1/\delta$ of length δ. Via the pointwise convergence above, get $n_0 < \infty$ large such that for all $n > n_0$

$$\mathbb{P}\left(\sup_{1\leqslant i\leqslant \frac{1}{\delta}} \left|\hat{q}_{\geqslant}^{(n)}(k,i\delta) - q_{\geqslant}^{(\infty)}(k,i\delta)\right| < \frac{\varepsilon}{4}\right) \geqslant 1-\varepsilon. \tag{3.72}$$

Write $G_n(\varepsilon,\delta)$ for the event in the above equation. Then on this event, by the choice of δ, for all i we have $|\hat{q}_{\geqslant}^{(n)}(k,i\delta) - \hat{q}_{\geqslant}^{(n)}(k,(i+1)\delta)| \leqslant \varepsilon/2$. Using monotonicity, for any $t \in [i\delta,(i+1)\delta]$, $|\hat{q}_{\geqslant}^{(n)}(k,i\delta) - \hat{q}_{\geqslant}^{(n)}(k,t)| \leqslant \varepsilon/2$. By the triangle inequality on $G_n(\varepsilon,\delta)$, for all $t \in [0,1]$ and $n > n_0$,

$$\begin{aligned}|\hat{q}_{\geqslant}^{(n)}(k,t) - q_{\geqslant}^{(\infty)}(k,t)| &\leqslant |\hat{q}_{\geqslant}^{(n)}(k,t) - q_{\geqslant}^{(n)}(k,i\delta)| + |\hat{q}_{\geqslant}^{(n)}(k,i\delta) - q_{\geqslant}^{(\infty)}(k,i\delta)| \\ &\quad + |q_{\geqslant}^{(\infty)}(k,i\delta) - q_{\geqslant}^{(\infty)}(k,t)| \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon.\end{aligned}$$

Since n_0 is independent of t, this completes the proof.

■

3.4 3.4.3. Proof of the tail exponent for the limiting degree distribution

The aim of this section is to prove the tail bound (3.8).

3.4 3.4.3.1. Overview of the proof

First note that the lower tail bound is obvious since with probability γ, $D_{\boldsymbol{\theta}}$ stochastically dominates D_α and by (3.2), D_α has the asserted tail behavior. The main crux is then proving the upper bound, namely

$$\mathbb{P}(D_{\boldsymbol{\theta}} \geqslant x) \leqslant c'/x^{2+\alpha}. \tag{3.73}$$

Recall that, from the characterization in the previous section 3.4.2.1, $D_{\boldsymbol{\theta}}$ can be separated into:

(a) with probability γ, $D_{\boldsymbol{\theta}}^{\mathsf{out}} = D_\alpha + N_\beta^{D_\alpha}[0, a]$;

(b) with probability $1 - \gamma$, $D_{\boldsymbol{\theta}}^{\mathsf{out}} = 1 + N_\beta[0, \mathsf{Age}]$.

As just mentioned, D_α already has the asserted tail behavior so we need to show that the components in this decomposition which are not D_α cannot contribute to the power-law tail. We will do this by proving that the distributions of $N_\beta^{D_\alpha}[0, a]$ and $N_\beta[0, \mathsf{Age}]$ both have exponential tails.

Note that both components $N_\beta^{D_\alpha}[0, a]$ and $N_\beta[0, \mathsf{Age}]$ correspond to the cumulative effect of evolution *after* the changepoint has occurred. Recalling the intuitive argument set out in Section 3.3.5, these components correspond to the growth occurring in the $O(1)$ amount of continuous time it takes the system to grow from size $n\gamma$ to n—proving that they also have an exponential tail is the final piece in showing that they cannot change the power-law exponent.

3.4 3.4.3.2. The upper bound

Recall Definition 3.4.10 of the Yule process and in particular Lemma 3.4.11 on the finite-time marginal distribution of the Yule process. Recall that in the description of the limit random variable $D_{\boldsymbol{\theta}}$, with probability $1 - \gamma$, $D_{\boldsymbol{\theta}} = N_\beta[0, \mathsf{Age}] \leqslant_{\mathsf{st}} N_\beta[0, a]$ where as before $\leqslant_{\mathsf{st}}$ represents stochastic domination. Now define

$$\nu = 2 + \beta, \qquad K = \lfloor 1 + \beta \rfloor \tag{3.74}$$

As before, let Y_ν^K be a rate ν Yule process started with K individuals at time zero. Comparing the rate of production of new individuals in the point process $\mathcal{P}_\beta$ with Y_ν^K, we get that $N_\beta[0, a] \leqslant_{\mathsf{st}} Y_\nu^K(a)$. By Lemma 3.4.11, $Y_\nu^K(a)$ is the sum of K independent Geometric

random variables. Using the fact that a geometric random variable has finite moment generating function in a neighborhood of zero and an elementary Chernoff bound implies that there exist constants $\kappa, \kappa' > 0$ such that for all $x \geqslant 1$, we have an exponential tail bound,

$$\mathbb{P}(N_\beta[0, \mathsf{Age}] > x) \leqslant \mathbb{P}(Y_\nu^K(a) > x) \leqslant \kappa' \exp(-\kappa x), \tag{3.75}$$

Thus when with probability $1 - \gamma$ $D_{\boldsymbol{\theta}} = N_\beta[0, \mathsf{Age}]$ then the corresponding random variable has exponential tail. Thus the main contribution to the tail arises when with probability γ, $D_{\boldsymbol{\theta}} = D_\alpha + N_\beta^{D_\alpha}[0, a]$. Arguing as above (and assuming $\beta \geqslant 1$), conditional on $D_\alpha = k$, we have

$$N_\beta^{D_\alpha}[0, a] \leqslant_{\mathsf{st}} \sum_{j=1}^{k} Y_\nu^{K,(j)}(a),$$

where, as in (3.63), $\left\{Y_\nu^{K,(j)}(\cdot) : j \geqslant 1\right\}$ are a collection of independent rate ν Yule processes each started at time zero with K individuals and independent of D_α. The following elementary lemma completes the proof.

Lemma 3.4.19. *Let $D \geqslant 1$ be non-negative integer valued random variable with $\mathbb{P}(D \geqslant x) \leqslant c/x^\gamma$ for all $x \geqslant 1$, for two constants $c, \gamma > 0$. Let $\{Y_i : i \geqslant 1\}$ be a sequence of independent and identically distributed positive integer valued random variables, independent of D. Consider the random variable $D^* := \sum_{j=1}^{D} Y_i$. If Y_1 has finite moment generating function in a neighborhood of zero then there exists a constant $c' > 0$ such that for all $x \geqslant 1$,*

$$\mathbb{P}(D^* \geqslant x) \leqslant c'/x^\gamma.$$

Proof: For the rest of the proof, write $\mu = \mathbb{E}(Y_1) < \infty$. Then note that

$$\begin{aligned}\mathbb{P}(D^* \geqslant x) &\leqslant \sum_{j=1}^{\frac{x}{2\mu}} \mathbb{P}(D = j)\, \mathbb{P}(\sum_{i=1}^{j} Y_i \geqslant x) + \mathbb{P}\left(D \geqslant \frac{x}{2\mu}\right), \\ &\leqslant \mathbb{P}(\sum_{i=1}^{\frac{x}{2\mu}} Y_i \geqslant x) + \frac{c}{x^\gamma},\end{aligned}$$

where the second equation follows using the fact that $Y_i \geqslant 1$ for all i and the tail bound for D from the hypothesis of the lemma. To complete the proof, note that standard large deviation bounds imply (since Y_i has a finite moment generating function about zero) imply that there exist constants κ, κ' such for all large x

$$\mathbb{P}\left(\sum_{i=1}^{\frac{x}{2\mu}} Y_i \geqslant x\right) \leqslant \kappa' \exp(-\kappa x).$$

This completes the proof. ■

The only item left to complete the proof of Theorem 3.2.1 is to show that the change point **does** change the degree distribution from the original (no change point) model. In Section 3.4.5 we will carry out a detailed analysis of the density of leaves which in particular will show that the asymptotic density of leaves $p_{\boldsymbol{\theta}}(1) \neq p_\alpha(1)$.

3.4 3.4.4. Analysis of the maximal degree

The aim of this section is to prove Theorem 3.2.2. First note that, for any fixed $k \geqslant 1$, writing $M_{\gamma n}(k)$ for the k-th maximal degree of a vertex in $\mathcal{T}_{\gamma n}$ namely in the tree just before the change point, using (3.3) implies that $M_{\gamma n}(k)/n^{1/(2+\alpha)}$ converges weakly to a strictly positive random variable. Since $M_n(k) \geqslant M_{\gamma n}(k)$, this implies that given any $\varepsilon > 0$ and any fixed $k \geqslant 1$, there exists a constant $K'_\varepsilon > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\frac{M_n(k)}{n^{1/(2+\alpha)}} > K'_\varepsilon\right) > 1 - \varepsilon.$$

Thus to complete the proof of Theorem 3.2.2 we need to show, given any $\varepsilon > 0$, there exists $K_\varepsilon < \infty$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{M_n(1)}{n^{1/(2+\alpha)}} < K_\varepsilon\right) \geqslant 1 - \varepsilon. \tag{3.76}$$

For any vertex $v \in [n]$ time point $m \in [n]$, write $\deg(v, m)$ for the degree of vertex v in $\mathcal{T}_m$ with the obvious convention that $\deg(v, k) = 0$ if $k < v$. Then note that $M_n(1) = \max(M_{\text{pre}}(n), M_{\text{post}}(n))$ where

$$M_{\text{pre}}(n) := \max_{v\in[1,n\gamma]} \deg(v, n), \qquad M_{\text{post}}(n) := \max_{v\in[n\gamma+1,n]} \deg(v, n). \tag{3.77}$$

Let us first analyze the maximal degree of vertices that appeared after the change point. Recall the constant a from (3.5) and ν, K from (3.74).

Lemma 3.4.20. *We have* $\mathbb{P}(M_{\text{post}}(n) > 2Ke^{\nu(a+1)} \log n) \to 0$ *as* $n \to \infty$.

Proof: We will assume $\beta \geqslant 1$ below. Else replace β with one in the rest of the argument below. For simplicity write $k_n = 2Ke^{\nu(a+1)} \log n$. Recall that in the continuous-time embedding, T_v represents the time of birth of vertex v and further for $v \in [\gamma n + 1, n]$, each such vertex is equipped with a offspring point process $\mathcal{P}^v_\beta$. As in Section 3.4.3, $1 + \mathcal{P}_\beta \leqslant_{\mathsf{st}} Y^K_\nu$ where Y^K_ν is a rate ν Yule process started with K individuals at time zero. Now note that via our continuous-time embedding,

$$M_{\text{post}}(n) := \max_{v\in[\gamma n+1,n]} (1 + \mathcal{P}^v_\beta(0, \tau_n - T_v)),$$

since by time τ_n, a vertex born after the change time has been alive for $\tau_n - T_v \leqslant \tau_n - \tau_{\gamma n} := \Upsilon_n$ units of time. Now

$$\begin{aligned}
\mathbb{P}(M_{\text{post}}(n) > k_n) &\leqslant \mathbb{P}(M_{\text{post}}(n) > k_n, \Upsilon_n < a + 1) + \mathbb{P}(\Upsilon_n > a + 1), \\
&\leqslant \mathbb{P}(\max_{v\in[\gamma n+1,n]} (1 + \mathcal{P}^v_\beta(0, a + 1)) > k_n) + \mathbb{P}(\Upsilon_n > a + 1).
\end{aligned} \tag{3.78}$$

Using Lemma 3.4.7 we have $\limsup_{n\to\infty} \mathbb{P}(\Upsilon_n > a+1) = 0$. Let $\left\{Y^K_{\nu,v} : v \in [\gamma n+1, n]\right\}$ be a family of independent rate ν Yule processes started with K individuals at time zero. Using Lemma 3.4.11 a simple union bound and the choice of k_n implies $\mathbb{P}(\max_{v\in[\gamma n+1,n]} Y^v_\beta(a+1) > k_n) \to 0$. ■

Thus the above lemma implies that the maximal degree amongst vertices that arrive after the change point is $O_P(\log n)$. To complete the proof of (3.76), it is enough to show that (3.76) holds with $M_n(1)$ replaced by $M_{\mathrm{pre}}(1)$. Thus fix $\varepsilon \in (0,1)$. Using Proposition 3.4.6 fix $A = A_\varepsilon$ such that

$$\limsup_{n\to\infty} \mathbb{P}(\tau_{\gamma n} - \frac{1}{2+\alpha}\log \gamma n > A) \leqslant \varepsilon/2. \tag{3.79}$$

Now consider the following process $\mathsf{BP}^n_{\boldsymbol{\theta},\star}$:

(a) Run the process BP_α until time $t_n(A) := \frac{1}{2+\alpha}\log \gamma n + A$.

(b) At this time: all vertices in $\mathsf{BP}_\alpha(t_n)$ switch to the dynamics with parameter β namely each vertex now reproduces at rate proportional to its out-degree $+ 1 + \beta$.

(c) Run this process for an additional $a+1$ units of time where a is as in (3.5).

Abusing notation, let $M^\star_{\mathrm{pre},A}(1)$ denote the maximal degree by time $t_n + a + 1$ of all vertices born *before* time t_n. We can obviously couple the original process $\mathsf{BP}^n_{\boldsymbol{\theta}}$ and $\mathsf{BP}^n_{\boldsymbol{\theta},\star}$ such that on the set $\left\{\tau_{\gamma n} - \frac{1}{2+\alpha}\log \gamma n \leqslant A, \Upsilon_n \leqslant a+1\right\}$ we have $M_{\mathrm{pre}}(1) \leqslant M^\star_{\mathrm{pre},A}(1)$.

Further note that for any fixed K we have

$$\begin{aligned}\mathbb{P}\left(M_{\mathrm{pre}}(1) > Kn^{1/(2+\alpha)}\right) \leqslant& \mathbb{P}\left(M_{\mathrm{pre}}(1) > Kn^{1/(2+\alpha)}, \Upsilon_n < a+1, \tau_{\gamma_n} < \frac{1}{2+\alpha}\log \gamma n + A\right)\\ &+ \mathbb{P}(\Upsilon_n > a+1) + \mathbb{P}(\tau_{\gamma_n} > \frac{1}{2+\alpha}\log \gamma n + A).\end{aligned}$$

First choosing A appropriately as in (3.79) and using Lemma 3.4.7 we get that for any fixed K,

$$\limsup_{n\to\infty} \mathbb{P}(M_{\text{pre}}(1) > Kn^{1/(2+\alpha)}) \leqslant \limsup_{n\to\infty} \mathbb{P}(M^{\star}_{\text{pre},A}(1) > Kn^{1/(2+\alpha)}) + \varepsilon/2.$$

The following lemma completes the proof of (3.76).

Lemma 3.4.21. *Fix $A > 0$. Given any $\varepsilon > 0$, we can choose $K = K(A, \varepsilon) < \infty$ such that*

$$\limsup_{n\to\infty} \mathbb{P}(M^{\star}_{\text{pre},A}(1) > Kn^{1/(2+\alpha)}) \leqslant \varepsilon.$$

Proof: First note that until time $t_n(A)$, the process $\mathsf{BP}^n_{\theta,\star}$ is a the continuous-time version of a (non-change point) preferential attachment model with attachment parameter α. This continuous-time embedding was used to derive asymptotics for the maximal degree in [12, 13]. In particular the bounds derived in these papers imply the following for a fixed A: Write $\tilde{M}_n(1)$ for the maximal degree exactly at time $t_n(A)$. Then there exists $L = L(A, \varepsilon) < \infty$ such that

$$\limsup_{n\to\infty} \mathbb{P}(\tilde{M}_n(1) > Ln^{1/(2+\alpha)}) \leqslant \varepsilon/2. \tag{3.80}$$

Now note that on the event $\left\{\tilde{M}_n(1) \leqslant Ln^{1/(2+\alpha)}\right\}$ at time $t_n + a + 1$, the degree of every fixed vertex in the system is stochastically dominated by a rate ν Yule process started with $Ln^{1/(2+\alpha)}$ vertices at time zero and run for time $a + 1$ where ν is as in (3.74). Write D_n for such a random variable and note that by the description of the dynamics of the Yule process and Lemma 3.4.11, we have that

$$D_n \stackrel{d}{=} \sum_{j=1}^{Ln^{1/(2+\alpha)}} Y_{\nu,j}(a+1), \tag{3.81}$$

where $\{Y_{\nu,j}(a+1) : j \geqslant 1\}$ are iid Geometric random variables with $p = e^{-\nu(a+1)}$. Further note that using Proposition 3.4.6 on the size of the branching process, we can choose C such

that

$$\limsup_{n\to\infty} \mathbb{P}(|\mathsf{BP}^n_{\boldsymbol{\theta},\star}(t_n)| > Cn) \leqslant \varepsilon/2. \tag{3.82}$$

Thus on the "good" event

$$\mathcal{G}_n := \left\{ |\mathsf{BP}^n_{\boldsymbol{\theta},\star}(t_n)| \leqslant Cn, \tilde{M}_n(1) \leqslant Ln^{1/(2+\alpha)} \right\},$$

we have that

$$M^\star_{\mathrm{pre},A}(1) \leqslant_{\mathsf{st}} \max_{1\leqslant v\leqslant Cn} D^v_n := \mathcal{M}_n$$

where $\{D^v_n : v \geqslant 1\}$ is an iid sequence with distribution (3.81). Note that $\mathbb{E}(Y_{\nu,i}(a+1)) = e^{\nu(a+1)}$. Let $K := 10Le^{\nu(a+1)}$. Then standard large deviations for the Geometric distribution implies that there exists a constant $C' > 0$ such that for all $n \geqslant 1$

$$\mathbb{P}(D_n \geqslant Kn^{1/(1+\alpha)}) \leqslant \exp(-C'n^{1/(1+\alpha)}).$$

Thus by the union bound,

$$\mathbb{P}(\mathcal{M}_n > Kn^{1/(1+\alpha)}) \leqslant Cn\exp(-C'n^{1/(1+\alpha)}) \to 0, \tag{3.83}$$

as $n \to \infty$. Finally,

$$\limsup_{n\to\infty} \mathbb{P}(M^\star_{\mathrm{pre},A}(1) > Kn^{1/(2+\alpha)}) \leqslant \limsup_{n\to\infty} \mathbb{P}(\mathcal{G}^c_n) + \limsup_{n\to\infty} \mathbb{P}(\mathcal{M}_n > Kn^{1/(2+\alpha)}) \leqslant \varepsilon,$$

using (3.80), (3.82) and (3.83). This completes the proof of the lemma and thus the analysis of the maximal degree asymptotics. ■

3.4 3.4.5. Analysis of the proportion of leaves

The aim of this section is to prove Theorem 3.2.3.

Theorem 3.2.3 relates the fluctuations of $N_n(1,m)$ (the number of leaves in the graph of size n at time m) around $ntp_t^{(\infty)}$ (the *limiting* number of leaves in the $n \to \infty$ limit). To accomplish this we will need two ingredients:

1. A functional central limit theorem for $N_n(1,m)$ around its *expectation* $\mathbb{E}\, N_n(1,m)$.

2. A strong, uniform bound on the error between the expectation $\mathbb{E}\, N_n(1,m)$ and the limit $ntp_t^{(\infty)}$.

We will start by proving the expectation error bounds in Section 3.4.5.1 using a simple recursion for $\mathbb{E}\, N_n(1,m)$. Then, we prove the functional central limit theorem in Section 3.4.5.2 by analyzing the martingale associated with $N_n(1,m)$.

We will need some extra notation for the following subsections. For the rest of the proof, to ease notation we will write $N_n(m) := N_n(1,m)$ for the number of leaves in $\mathcal{T}_m$ and let $\hat{N}_n(t) = N_n(nt)$. Recall the asserted limiting proportion $\left\{p_t^{(\infty)} : 0 \leqslant t \leqslant 1\right\}$ from (3.10). For each $n \geqslant 2$ define the collection of real numbers $\mathbf{w}_n = \{w_m : 2 \leqslant m \leqslant n-1\}$

$$w_m = \begin{cases} \left(1 - \frac{1+\alpha}{(2+\alpha)m-1}\right) & \text{if } 2 \leqslant m \leqslant n\gamma - 1, \\ \left(1 - \frac{1+\beta}{(2+\beta)m-1}\right) & \text{if } n\gamma \leqslant m \leqslant n-1. \end{cases} \tag{3.84}$$

3.4 3.4.5.1. Expectation error bounds

The following proposition is the main result of this section.

Proposition 3.4.22. *There exists a constant $C < \infty$ independent of n such that the expectations satisfy*

$$\sup_{n \geqslant 1} \sup_{0 \leqslant t \leqslant 1} \left| \mathbb{E}(\hat{N}_n(t)) - ntp_t^{(\infty)} \right| \leqslant C. \tag{3.85}$$

Remark 8. Note that by Proposition 3.4.18, we know there exists a function $p^{(\infty)}(0,\cdot)$ such that $\hat{p}^{(n)}(0,t) \to p^{(\infty)}(0,t)$ for $0 < t \leqslant 1$. By the bounded convergence theorem, $\mathbb{E}(\hat{p}^{(n)}(0,t)) \to p^{(\infty)}(0,t)$. Thus the above proposition implies that $p^{(\infty)}(0,t) = p_t^{(\infty)}$. In

particular it shows that the degree distribution owing to the change point is **different** from the degree distribution without change point. This is the final nail in proving Theorem 3.2.1.

Remark 9. A similar result was shown in the context of no change point in [106, Section 8.6] and [46] (not just for leaves but for all fixed $k \geqslant 1$). Our proof uses slightly different ideas starting from the same point as in [106]. While we do not consider higher degree vertices, as in [106], the result above can be used as a building block to show identical error bounds for expectations of the number of higher degree vertices about limit constants.

Proof: To ease notation write $\vartheta_n(m) = \mathbb{E}(N_n(m))$. The main crux of the proof is studying a recursion relation for $\vartheta_n(m+1)$ in terms of $\vartheta_n(m)$. We will give a careful analysis of the time period before the change point and then describe how the same ideas give the result for after the change point.

For each $1 < m \leqslant n$ write $\mathcal{L}_{m+1}$ for the event that vertex $m+1$ connects to a leaf vertex in $\mathcal{T}_m$. Then note that conditioning on $\mathcal{T}_m$, when $m < n\gamma$ we have

$$\begin{aligned}\mathbb{E}(N_n(m+1)|\mathcal{T}_m) &= N_n(m) + 1 - \mathbb{P}(\mathcal{L}_{m+1}|\mathcal{T}_m) \\ &= N_n(m) + 1 - \frac{(1+\alpha)N_n(m)}{(2+\alpha)m - 1} \end{aligned} \tag{3.86}$$

When $m \geqslant n\gamma$ we have the same recursion as above but with α replaced by β. Taking full expectations and simplifying gives the following recursion:

$$N_n(m+1) = 1 + w_m N_n(m), \qquad \vartheta_n(m+1) = 1 + w_m \vartheta_n(m), \tag{3.87}$$

where $\{w_m : 2 \leqslant m \leqslant n\}$ are as defined in (3.84).

Before the change point: Repeatedly using this recursion and using the boundary condition $\vartheta_n(2) = 1$ gives for $m + 1 \leqslant n\gamma$,

$$\vartheta_n(m+1) = \sum_{s=2}^{m} \prod_{k=s}^{m} \left(1 - \frac{(1+\alpha)}{(2+\alpha)k - 1}\right) \tag{3.88}$$

Now fix $s_0 \geqslant 1$ large enough such that the following three conditions hold:

(i) For all $k \geqslant s_0$

$$\log k + \gamma \leqslant \sum_{i=1}^{k} \frac{1}{i} \leqslant (\log k + \gamma) + \frac{1}{k}.$$

Here γ is the Euler-Mascheroni constant. See [17].

(ii) For all $k \geqslant s_0$, $1 - \frac{(1+\alpha)}{(2+\alpha)k-1} \geqslant 1/2$.

(iii) We may choose a constant $C < \infty$ such that for all $k \geqslant 1$,

$$\sum_{i=k}^{\infty} \frac{1}{((2+\alpha)k-1)^2} \leqslant \frac{C}{k}. \tag{3.89}$$

Further there exists a constant C' such that for all $s > s_0$, $|\exp(C/s) - 1| \leqslant C'/s$ and

$$\left| \left(1 - \frac{(1+\alpha)}{(2+\alpha)s-1}\right) - e^{-\frac{(1+\alpha)}{(2+\alpha)s-1}} \right| \leqslant \frac{C'}{s^2}.$$

To ease notation, for the rest of the proof let $\delta = (1+\alpha)/(2+\alpha)$. Using the elementary inequality $1 - x \leqslant e^{-x}$ for $x \in (0,1)$ and the choice of s_0 above, the following inequalities with a constant $C = C(s_0, \alpha) < \infty$ are readily verified:

(A) For all $m \geqslant s \geqslant s_0$,

$$\left| e^{-\sum_{i=s}^{m} \frac{\delta}{i}} - \left(\frac{s}{m}\right)^{\delta} \right| \leqslant C \frac{s^{\delta-1}}{m^{\delta}}. \tag{3.90}$$

(B) For all $m \geqslant s \geqslant s_0$,

$$\left| e^{-\sum_{i=s}^{m} \frac{(1+\alpha)}{(2+\alpha)i}} - e^{-\sum_{i=s}^{m} \frac{(1+\alpha)}{(2+\alpha)i-1}} \right| \leqslant C \frac{s^{\delta-1}}{m^{\delta}}. \tag{3.91}$$

(C) For all $m \geqslant s \geqslant s_0$,

$$\prod_{k=s}^{m} \left(1 - \frac{(1+\alpha)}{(2+\alpha)k-1}\right) \leqslant C \left(\frac{s}{m}\right)^{\delta}. \tag{3.92}$$

Now note that by the "Lindeberg" trick, for any $s \leqslant m$ and two collections of non-negative numbers $\{w_k : s \leqslant k \leqslant m\}$ and $\{z_k : s \leqslant k \leqslant m\}$ we have

$$\left|\prod_{k=s}^{m} w_k - \prod_{k=s}^{m} z_k\right| \leqslant \sum_{k=s}^{m} |w_k - z_k| \prod_{s \leqslant l < k} z_k \prod_{l > k} w_k \tag{3.93}$$

Using this with $w_k = 1 - \frac{(1+\alpha)}{(2+\alpha)k-1}$ and $z_k = e^{-\frac{(1+\alpha)}{(2+\alpha)k-1}}$ and using (3.90), (3.91) and (3.92) gives the following lemma.

Lemma 3.4.23. *Fix s_0 as above. Writing $\delta = (1+\alpha)/(2+\alpha)$ there exists a constant $C < \infty$ such that for all $m \geqslant s \geqslant s_0$,*

$$\left|\prod_{k=s}^{m} \left(1 - \frac{(1+\alpha)}{(2+\alpha)k-1}\right) - \left(\frac{s}{m}\right)^{\delta}\right| \leqslant C\frac{s^{\delta-1}}{m^{\delta}}.$$

Now using the form of the expectation $\vartheta_n(m)$ in (3.88), the error bound in the above lemma and the integral comparison

$$\frac{1}{m^{\delta}} \int_{s_0}^{m-1} x^{\delta} dx \leqslant \sum_{s_0+1}^{m} \left(\frac{s}{m}\right)^{\delta} \leqslant \frac{1}{m^{\delta}} \int_{s_0+2}^{m+1} x^{\delta} dx,$$

shows that there exists a constant C such that for $m \leqslant n\gamma$

$$|\vartheta_n(m) - \frac{m}{\delta}| \leqslant C. \tag{3.94}$$

This is the assertion for the expected number of leaves before the change point.

After the change point: We now describe the evolution of $\vartheta_n(m)$ for $n\gamma < m \leqslant n$. We only give the basic idea as the details are the same as before the change point. First note that by the above analysis, there exists a constant C such that $|\vartheta_n(n\gamma) - n\gamma/\delta| \leqslant C$. Now the evolution of the process after γn is as in (3.86) with α replaced by β. Thus starting at

$m > n\gamma$ and using the argument above we get

$$\vartheta_n(m+1) := \sum_{s=n\gamma+1}^{m} \prod_{j=s}^{m} \left(1 - \frac{1+\beta}{(2+\beta)j-1}\right) + \vartheta_n(n\gamma) \prod_{j=n\gamma}^{m} \left(1 - \frac{1+\beta}{(2+\beta)j-1}\right) \tag{3.95}$$

Simplifying notation and writing $m = nt$ where $\gamma \leqslant t \leqslant 1$ and repeating the arguments above it is easy to check that there exists a constant C independent of n such that

$$|\vartheta_n(nt) - ntp_t^{(\infty)}| \leqslant C, \tag{3.96}$$

where $p_t^{(\infty)}$ is as in (3.10). This completes the proof.

■

3.4 3.4.5.2. Proof of Theorem 3.2.3

A central limit theorem for the number of leaves $N_n(n)$ (in fact all degree counts $N_n(k,n)$) at time n in the setting of no change point was established in [91]. We will extend this to a functional central limit theorem in the change point setting. First recall the function δ_α from (3.11). Define the stochastic process

$$M_n^*(t) = \begin{cases} t^{\delta_\alpha} \dfrac{(N_n(nt) - \vartheta_n(nt))}{\sqrt{n}} & \text{if } t \leqslant \gamma \\ \gamma^{\delta_\alpha} \left(\dfrac{t}{\gamma}\right)^{\delta_\beta} \dfrac{(N_n(nt) - \vartheta_n(nt))}{\sqrt{n}} & \text{if } t \geqslant \gamma \end{cases} \tag{3.97}$$

Recall the process $M(\cdot)$ in (3.15) and the relationship between M and G. Using Proposition 3.4.22 and the continuous mapping theorem, it is enough to show the following result.

Proposition 3.4.24. *We have $M_n^*(\cdot) \xrightarrow{\text{w}} M(\cdot)$ on $D[0,1]$ as $n \to \infty$.*

Proof: The main idea is to study martingales associated with the $\{N_n(m) : 2 \leqslant m \leqslant n\}$ and then use the martingale functional central limit theorem. There are an enormous number of variants of such functional limit theorems under a multitude of conditions. We quote the

specific form relevant to this setting. Recall the function $\phi(\cdot)$ and the corresponding diffusion $M(\cdot)$ defined in (3.16).

Theorem 3.4.25. *[[47, 49]] For each $n \geqslant 1$, let $\{M_n(m) : 1 \leqslant m \leqslant n\}$ be a mean zero martingale with finite second moments adapted to a filtration $\{\mathcal{F}_n(m) : 1 \leqslant m \leqslant n\}$. Write $\{X_n(m) : 1 \leqslant m \leqslant n\}$ for the associated martingale difference sequence namely $X_n(m) = M_n(m) - M_n(m-1)$ with $M_n(0) = 0$. Assume the following two hypotheses:*

(i) *For each $0 \leqslant t \leqslant 1$*

$$V_n(nt) := \sum_{m=1}^{nt} \mathbb{E}\left([X_n(m)]^2 \middle| \mathcal{F}_n(m-1) \right) \xrightarrow{\text{P}} \phi(t), \qquad \text{as } n \to \infty. \tag{3.98}$$

(ii) *For each fixed $\varepsilon > 0$*

$$\sum_{m \leqslant n} \mathbb{E}\left([X_n(m)]^2 \mathbb{1}\left\{ |X_n(m)| > \varepsilon \right\} \middle| \mathcal{F}_n(m-1) \right) \xrightarrow{\text{P}} 0. \tag{3.99}$$

Then defining the process $\bar{M}_n(t) := M_n(nt)$, one has $\bar{M}_n \xrightarrow{\text{w}} M$ in $D[0,1]$.

For our example (following [91]) define the process

$$N_n^*(m) = \frac{N_n(m) - \vartheta_n(m)}{\prod_{j=2}^{m-1} w_j}, \qquad 2 \leqslant m \leqslant n. \tag{3.100}$$

Here w_j is as in (3.84). Using the recursion (3.87) results in the following lemma.

Lemma 3.4.26. *The process N_n^* is a martingale with respect to the filtration generated by $\{\mathcal{T}_m : 2 \leqslant m \leqslant n\}$.*

Now define the corresponding martingale differences $d_n(m) = N_n^*(m) - N_n^*(m-1)$. Define $\Delta_n(m) = \mathbb{1}\{m+1 \text{ connects to a non-leaf vertex in } \mathcal{T}_{m-1}\}$. Then simple algebra and (3.87) implies that for $m \leqslant n\gamma$

$$d_n(m) = \frac{1}{\prod_{j=2}^{m-1} w_j} \left[\Delta_n(m) + N_n(m-1) \frac{(1+\alpha)N_n(m-1)}{(2+\alpha)(m-1) - 1} - 1 \right], \tag{3.101}$$

and

$$\mathbb{E}(\Delta_n(m)|\mathcal{T}_{m-1}) = 1 - \frac{(1+\alpha)N_n(m-1)}{(2+\alpha)(m-1)-1} \tag{3.102}$$

For $m \geqslant n\gamma$ we have identical formulae as (3.101) and (3.102) but now α is replaced by β. For the rest of the argument we will replace the denominator for the second term $(2+\alpha)(m-1)-1$ by $(2+\alpha)(m-1)-1$. It is easy to check that the error is negligible and will ease presentation.

Now use Proposition 3.4.18 which allows us to uniformly approximate $N_n(m-1)/(m-1)$ by $p^{(\infty)}_{m/n}$. Further the asymptotics of $\prod_{j=2}^m w_j$ derived in the previous section implies that for $m \leqslant n\gamma$, $\prod_{j=2}^m w_j \sim m^{-\delta_\alpha}$ while for $m > n\gamma$, $\prod_{j=2}^m w_j \sim (n\gamma)^{-\delta_\alpha}(m/n\gamma)^{-\delta_\beta}$ where $\delta_\alpha, \delta_\beta$ as defined in (3.11). Taking conditional expectations in (3.101), using (3.102) and using the above approximations results in

$$\mathbb{E}([d_n(m)]^2|\mathcal{T}_{m-1}) \sim \begin{cases} m^{2\delta_\alpha}\left[\delta_\alpha p^{(\infty)}_{m/n}(1-\delta_\alpha p^{(\infty)}_{m/n})\right] & \text{if } m \leqslant n\gamma, \\ \\ (n\gamma)^{2\delta_\alpha}\left(\frac{m}{n}\right)^{2\delta_\beta}\left[\delta_\beta p^{(\infty)}_{m/n}(1-\delta_\beta p^{(\infty)}_{m/n})\right] & \text{if } m \geqslant n\gamma \end{cases} \tag{3.103}$$

Now consider the martingale

$$M_n(m) := \frac{1}{n^{\delta_\alpha+1/2}} \frac{N_n(m) - \vartheta_n(m)}{\prod_{j=2}^{m-1} w_j}, \qquad 2 \leqslant m \leqslant n. \tag{3.104}$$

We will apply Theorem 3.4.25 to this martingale. Let $\{X_n(m) : 2 \leqslant m \leqslant n\}$ denote the corresponding martingale differences. First fix $t \leqslant \gamma$ and recall the definition of the cumulative conditional variance $V_n(nt)$ until time t in (3.98). Using the first expression in (3.103) we get

$$\begin{aligned} V_n(nt) &\sim \frac{1}{n^{2\delta_\alpha+1}} \sum_{j=1}^{nt} j^{2\delta_\alpha}\left[\delta_\alpha p^{(\infty)}_{j/n}(1-\delta_\alpha p^{(\infty)}_{j/n})\right] \\ &\to \int_0^t s^{2\delta_\alpha}\left[\delta_\alpha p^{(\infty)}_s(1-\delta_\alpha p^{(\infty)}_s)\right] ds = \phi(t), \end{aligned}$$

as $n \to \infty$. Thus (3.98) is satisfied for $t \leqslant \gamma$. A similar calculation now incorporating the second expression in (3.103) implies that (3.98) is satisfied for all $t \in [0,1]$ with ϕ as in (3.16).

Now let us check the second condition namely (3.99). Note that for $m \leqslant n\gamma$, $X_n(m) \geqslant \varepsilon$ implies that $3m^{\delta_\alpha} \geqslant \varepsilon n^{\delta_\alpha + 1/2}$. For large n this is impossible for all $m \leqslant n\gamma$. A similar calculation for $m > n\gamma$ completes the proof of (3.99). Using Theorem 3.4.25 we get that $M_n(n\cdot) \xrightarrow{\text{w}} M(\cdot)$ in $D[0,1]$. Using the asymptotics for $\prod_{j=2}^{m} w_j$ derived in Section 3.4.5.1, Lemma 3.4.23 now completes the proof of Proposition 3.4.24 and thus Theorem 3.2.3. ■

3.4 3.4.6. Consistency of the estimator

The aim of this section is to prove Theorem 3.2.4. Fix a truncation level $\varepsilon > 0$ from zero as in the theorem. Recall the time-averaged proportion of leaves before and after each time t namely (3.22) and (3.21). Also recall the expression for the limiting proportion of leaves from (3.10). For any fixed interval $[s,t] \subseteq [0,1]$, define $H[s,t]$ by

$$H[s,t] := \frac{1}{t-s} \int_s^t p_u^{(\infty)} du. \tag{3.105}$$

The interpretation is as follows: the above gives the expected proportion of leaves in the large network limit if one were to sample a time point $U \in [s,t]$ uniformly at random. Now define the two functions ${}_t h^{(\infty)}$ and $h_t^{(\infty)}$ via:

(a) **Case 1:** For $\varepsilon \leqslant t \leqslant \gamma$

$${}_t h^{(\infty)} := p_\gamma^{(\infty)}, \qquad h_t^{(\infty)} := \frac{\gamma - t}{1-t} p_\gamma^{(\infty)} + \frac{1-\gamma}{1-t} H[\gamma, 1]$$

(b) **Case 2:** For $t > \gamma$

$${}_t h := \frac{\gamma - \varepsilon}{t - \varepsilon} p_\gamma + \frac{t-\gamma}{t-\varepsilon} H([\gamma, t]), \qquad h_t := H([t,1]).$$

In a similar vein to (4.2), define the function

$$D(t) := (1-t)|_t h^{(\infty)} - h_t^{(\infty)}|, \qquad t \in [\varepsilon, 1]. \tag{3.106}$$

Routine algebra shows that

$$D(t) := \begin{cases} (1-\gamma)\left|p_\gamma^{(\infty)} - H[\gamma,1]\right| & \text{for } \varepsilon \leqslant t \leqslant \gamma. \\ \\ (1-\varepsilon)\left|H[\varepsilon,t] - H[\varepsilon,1]\right| & \text{for } t > \gamma. \end{cases} \tag{3.107}$$

Using the form of the limit proportion $p_t^{(\infty)}$ from (3.10) the following result is easy to check.

Lemma 3.4.27. *Fix $\varepsilon < \gamma$ and assume $\alpha \neq \beta$. Then $D(\cdot)$ is a continuous function on $[\varepsilon, 1]$ such that $D(\cdot)$ is constant on the interval $[\varepsilon, \gamma]$ and then is strictly monotonically decreasing on the interval $[\gamma, 1]$ with $D(t) \to 0$ as $t \to 1$. Further the function has a strictly negative right derivative at γ namely*

$$\partial_+ D(\gamma) := \lim_{t \downarrow \gamma} \frac{D(t) - D(\gamma)}{t - \gamma} < 0. \tag{3.108}$$

Now Theorem 3.2.3 immediately results in the following result.

Lemma 3.4.28. *Fix $\varepsilon > 0$. Then*

$$\sup_{t \in [\varepsilon, 1]} |D_n(t) - D(t)| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Combining Lemmas 3.4.27 and 3.4.28 completes the proof. ■

CHAPTER 4

Changepoint: simulations and analysis of real data

4.1 4.1. Introduction

New random graph models are proposed almost every day, but in the world of time-evolving, growing networks, preferential attachment still is king. The mechanism of preferential attachment is simple, mathematically rich, and also generates graphs with some real-world characteristics. However, the preferential attachment paradigm still falls short of being able to stand as a convincing model for many real-world, temporal networks.

To be more specific, many real-world networks resemble preferential attachment graphs at time $t = 1$ (i.e. in the aggregate view), but their evolution *from* $t = 0$ until $t = 1$ doesn't resemble preferential attachment at all. Therefore, we see that in order to build upon and extend the preferential attachment model to real-world analysis, it is not only helpful but *necessary* to start looking at the entire history of the graph.

The goal of this chapter is to further evaluate our approach to change point detection on simulated preferential attachment data, but also to examine to what extent two real-world temporal networks resemble preferential attachment graphs and how we can apply some of our change point insights to them.

In Section 4.2 we discuss some finer, practical points relating to the change point estimator proposed in Chapter 3 and evaluate its performance on simulated preferential attachment graphs. In Section 4.3 we investigate these Chapter 3-inspired graph functions on some real, temporal network data built from arXiv and CourtListener citation data to see what we can uncover.

4.2 4.2. Change point: further notes and simulations

4.2 4.2.1. Preferential attachment: the role of functions

We begin with the case that our change point estimator was designed for: linear preferential attachment trees. As these graphs have already been discussed at length in Chapter 3, let us dive straight into the details.

Linear preferential trees have been studied exhaustively. The interesting bits for us lie in comparing the trees *with* vs. *without* change point. In light of our findings from Chapter 3, the main takeaway from these simulations is that the effect of a change point cannot really be felt in the aggregate statistics calculated from the graph. We've already seen that the degree exponent does not respond to change point.

And for many other statistics that do respond measurably to the addition of a change point, there is no way to tell from the statistic whether the value of the statistic is due to a graph with change point or whether it is due to a graph without changepoint, but with a slightly different attachment parameter value. For example, the proportion of leaves from a non-change point model with attachment parameter α_0 can be easily generated from a change point model with different α_1 and β by solving equation 3.10.

The key function for the change point analysis is the leaf count over the history of the graph. Our estimator is a simple argmax of a function of the leaf count, so in theory we ought to be able to visually detect a change point from the path of this function alone. However, it is difficult to see the break in the raw count since the graph is continually growing. Normalizing by the size of the graph and plotting the leaf proportion gives the clearer picture seen in Figure 4.1.

Obviously, whether or not $\alpha > \beta$ holds determines the exact directionality of the kink. But that aside, the big picture is that we want to identify the time corresponding to the kink. Our strategy is to use $D_n(t)$:

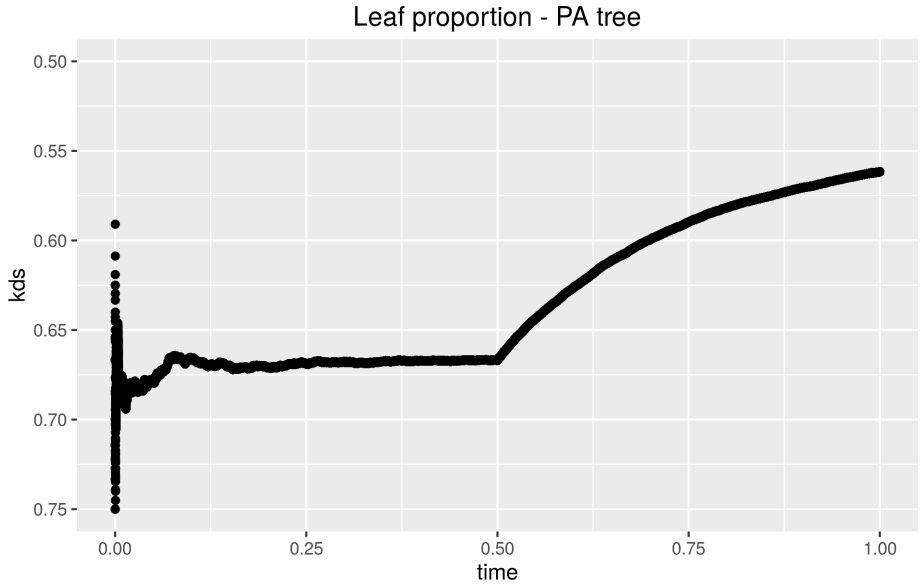


Figure 4.1: The proportion of leaves in a preferential attachment tree with $\gamma = 0.5$, $\alpha = 0$ and $\beta = 10$.

$$D_n(t) := (1-t)|{}_t h^{(n)} - h_t^{(n)}|, \qquad t \in [\varepsilon, 1]. \tag{4.1}$$

Where ${}_t h^{(n)}$ is the average proportion of leaves at each time point up until time t and $h_t^{(n)}$ is the average proportion of leaves at each time point after time t. Also recall that our estimator is essentially defined as:

$$\hat{\gamma} = \text{the last time that } D_n(t) \text{ was within } \frac{\omega(n)}{\sqrt{n}} \text{ of } \max_{t \in [\epsilon, 1]} D_n(t)$$

For a typical preferential tree with and without change point, the path of this function is straightforward to interpret: it's the scaled difference between the average proportion of leaves before and after each time t. Without a change point, this function is roughly constant at zero, since without a change point the proportion of leaves converges to a constant. With change point, the function is constant up until the change point and then decreases to 0 (see Figure 4.3).

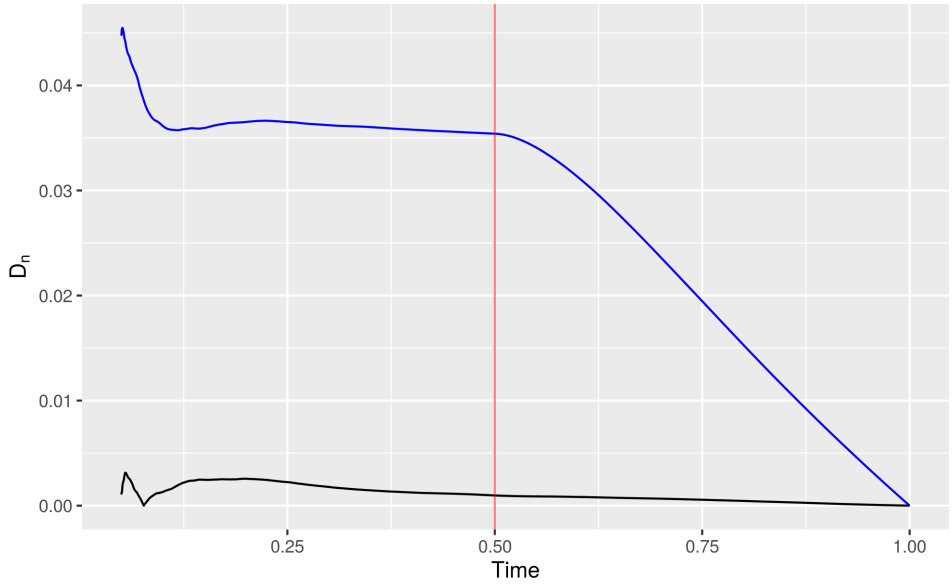


Figure 4.2: Plot of $D_n(t)$ for the with-changepoint model (blue) versus the no-changepoint model (black)

4.2 4.2.2. The behavior of $\hat{\gamma}$ in simulations

At this point, let's examine the behavior of $\hat{\gamma}$ in a little more detail.

4.2 4.2.2.1. The bias-variance tradeoff in ϵ

In the big scheme, our estimator has two main tuning parameters. It turns out that the most important is the ε truncation of the sequence $D_n(t)$ near zero. This truncation was introduced as a technical compensation for a divide-by-zero problem in the definition of the average proportion of leaves. If all we are interested in is consistency, then strictly speaking this is the only reason for this truncation.

However, for fixed sample sizes ε plays another role. Obviously, as the graph size $n \to \infty$ the proportion of leaves converges to a constant. But when the graph is small, the variance is extremely high. This translates to large fluctuations in $D_n(t)$ for t close to 0.

Even for graphs of, say, $n = 500,000$ vertices, there is some positive probability that the fluctuation is so large that even $\max_{t\in[\varepsilon,1]} D_n(t) \pm C \cdot \omega(n)/\log(n)$ fails to capture the

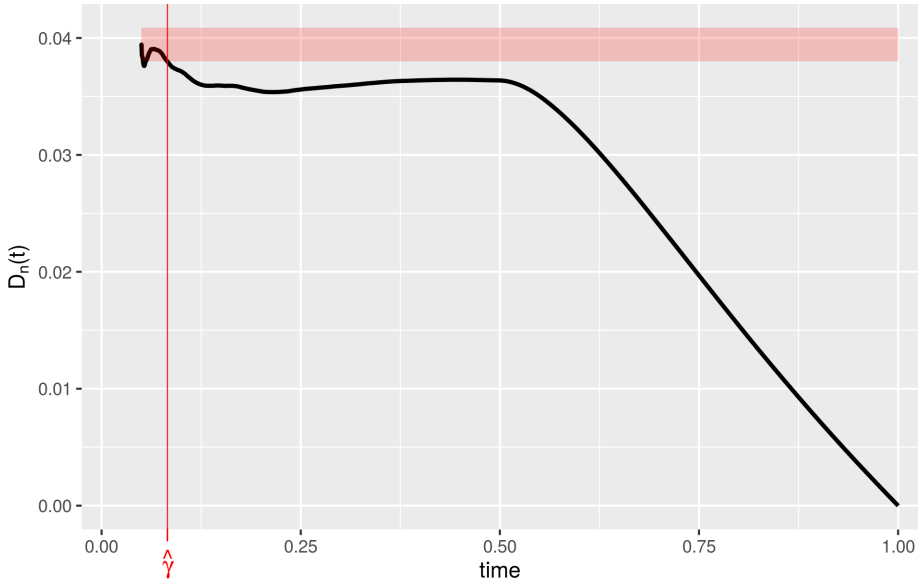


Figure 4.3: Illustration of when the $\log(n)/\sqrt{n}$ threshold (indicated by pink box) fails to include the true changepoint $\gamma = 0.5$.

main part of the $D_n(t)$ sequence, biasing the estimator *earlier* in time[1]. Furthermore, this is possible regardless of the values of the other parameters in the model since the small-t volatility is always present.

Therefore the practical effect of increasing ε is to reduce the variance in the estimator, especially when γ is close to 1. The downside is that automatically biases the estimator for any $\gamma < \varepsilon$, but for many real use cases the possibility of $\gamma < \varepsilon$ can be reasonably excluded.

As the reader will soon note, this ε-truncation plays an indirect role in much of the following discussion.

For all the later simulations unless otherwise specified, we will set $\varepsilon = 0.05$.

4.2 4.2.2.2. The bias-variance tradeoff in ω

The second important parameter in our setup is the choice of ω as defined in:

$$\mathcal{M}_n := \left\{ t \in [\varepsilon, 1] : |D_n(t) - \max_{t \in [\varepsilon,1]} D_n(t)| \leqslant \frac{\omega(n)}{\sqrt{n}} \right\}$$

[1]The scaling C is a small technical compensation which will be explained in the next subsection.

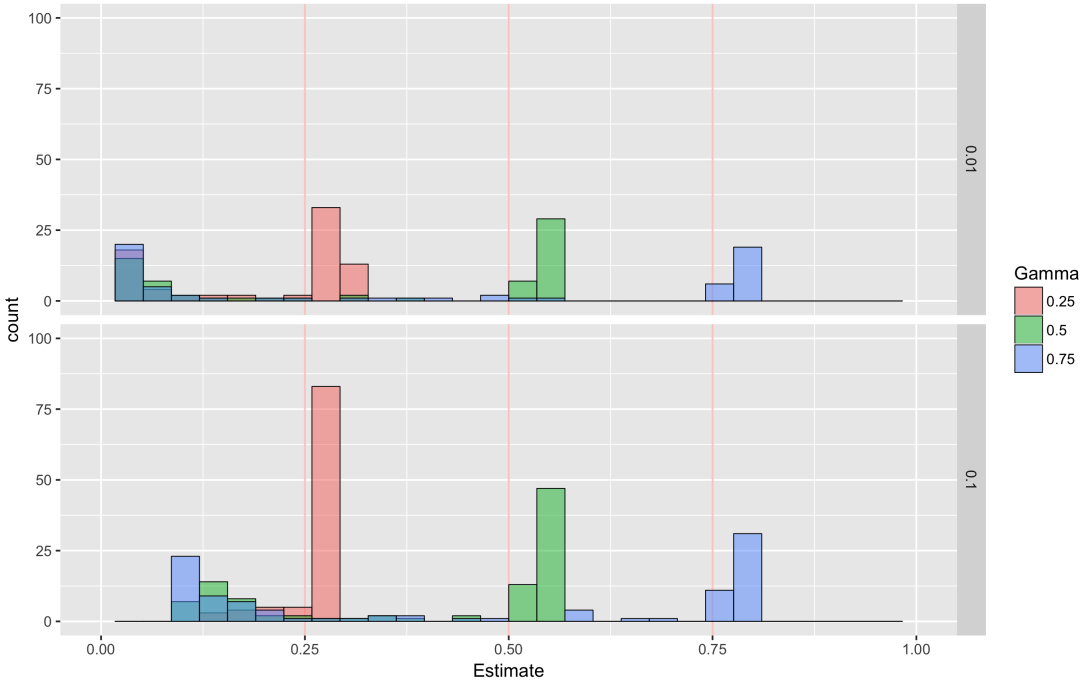


Figure 4.4: The effect of increasing ϵ from 0.01 to 0.10 for $\gamma \in \{0.25, 0.50, 0.75\}$. $\omega = \log$ in both cases.

$\hat{\gamma}$ is defined as the latest time $t \in \mathcal{M}_n$. In plain English, $\omega(n)/\sqrt{n}$ sets the fatness of the "tube" around $D_n(t)$ which defines the amount of fluctuations of $D_n(t)$ we can tolerate before declaring a change.

In the previous chapter we presented $\omega = \log$, but also remarked that this choice is arbitrary; essentially, consistency is still guaranteed so long as $\omega = o(\sqrt{n})$. To break down why this is, let $D(t)$ be the limit of $D_n(t)$ as $n \to \infty$. The functional central limit theorem 3.2.3 tells us that, for fixed $\varepsilon > 0$,

$$\sup_{t \in [\varepsilon, 1]} |D_n(t) - D(t)| = O_P\left(\frac{1}{\sqrt{n}}\right)$$

One way of looking at it is that ω is necessary to overcome the constant factor in the $O_P(\cdot)$ statement. Regardless of the interpretation, the main takeaway is that the threshold $\omega/\sqrt{n}$ was formulated as such purely to guarantee consistency. It makes no promises about

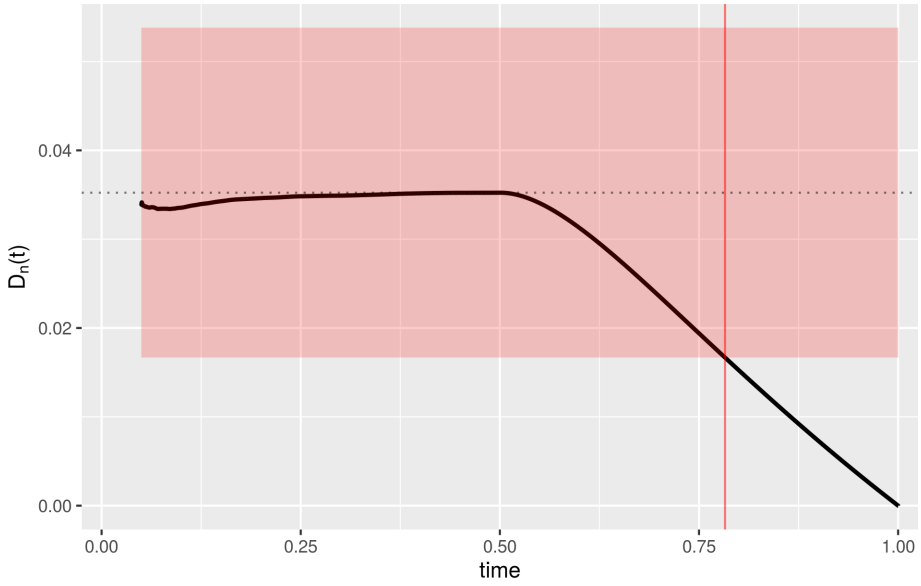


Figure 4.5: Effect of not normalizing $D_n(t)$ by $\max_t D_n(t)$ on a graph with $\gamma = 0.5$, $\alpha = 0$, $\beta = 10$, and $N = 500,000$. The red tube is the unnormalized threshold with vertical line at the estimate $\hat{\gamma}$.

small- or finite-sample performance. Indeed, if we inspect the threshold with $\omega(n) = \log(n)$ (Figure 4.5), then we can clearly see why the estimator can perform poorly.

The bias is very high for this estimator, simply because the threshold $\omega/\sqrt{n}$ is very large compared with the natural scale of D_n. This forces the set $\mathcal{M}_n$ to always include a significant portion of time after the break in the path of D_n. This bias is borne out in repeated simulation. In Figure 4.5, the threshold $\log(n)/\sqrt{n} \approx 0.019$ but the scale of the function $D_n(t)$ hovers around 0.035. Although as $n \to \infty$ this gap will close, the resulting bias even for 500,000 vertices is quite high.

This suggests an easy fix: simply scale the $\omega(n)/\sqrt{n}$ threshold by the natural scale of $D_n(t)$. Namely, adjust the definition of $\mathcal{M}_n$ like so:

$$\tilde{\mathcal{M}}_n := \left\{ t \in [\varepsilon, 1] : |D_n(t) - \max_{t \in [\varepsilon,1]} D_n(t)| \leqslant \left(\frac{\omega(n)}{\sqrt{n}} \cdot \max_{t \in [\varepsilon,1]} D_n(t) \right) \right\}$$

In much the same way as decreasing ε may cause variance to increase because the set $\mathcal{M}_n$ generally misses a larger portion of the constant phase of $D_n(t)$, scaling by $\max_{t \in [\varepsilon,1]} D_n(t)$

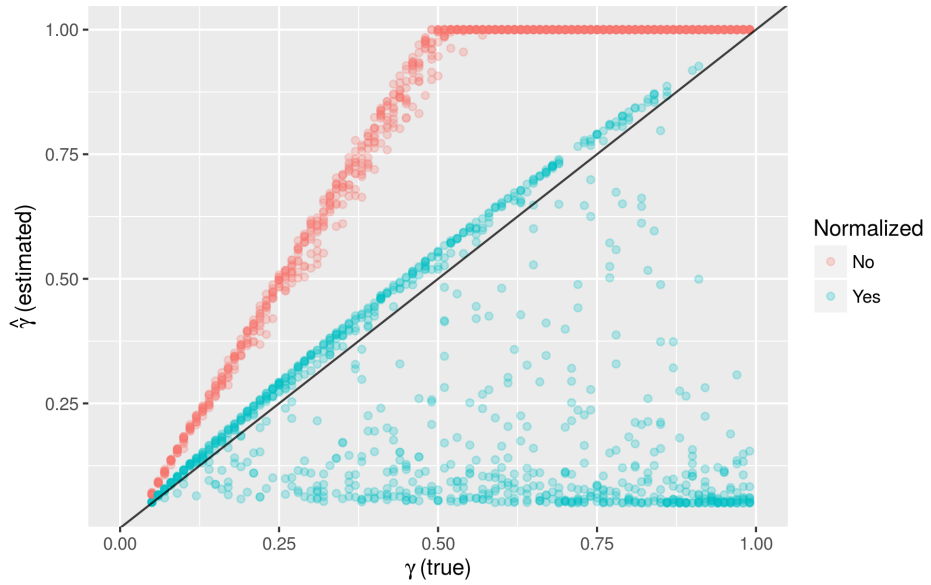


Figure 4.6: Normalized vs. unnormalized estimates for a change point of $\alpha = 0$ to $\beta = 10$ at various values of γ.

also injects more variance into our estimator. However, the reduction in bias is significant, see e.g. Figure 4.6. Note the somewhat similar end effect of removing the $(1 - t)$ scaling illustrated in Figure 3.4.

From this we can always see that the question of choice of ω boils down to essentially the same dynamic. For any fixed n, an ω going to ∞ relatively slowly will exhibit more variance but less bias compared to an ω going to ∞ relatively quickly.

It's worth noting though that the nature of the bias is slightly different than the bias resulting from a small ε. With bias due to small ε, the width of the tube threshold remains the same but the reference point $\max_t D_n(t)$ catches a point much earlier in the history of the graph, generally resulting in a *backwards* (towards earlier times) shift in $\hat{\gamma}$s. With bias due to large or unscaled $\omega(n)/\log(n)$ however, $\hat{\gamma}$ is clearly shifted *forwards* in time. Another look back at 4.4 versus 4.6 reveals this clearly.

From now on, unless otherwise specified the estimates shown and discussed will always be this scaled version of $\hat{\gamma}$, calculated using $\tilde{\mathcal{M}}_n$ shown above.

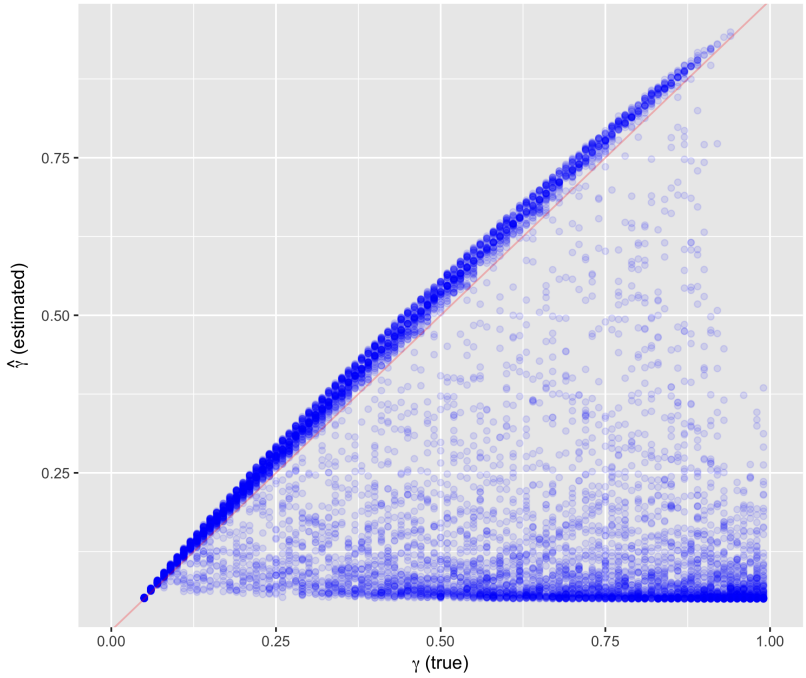


Figure 4.7: γ vs. $\hat{\gamma}$ for various $\gamma \in [0.05, 1.00]$. Changepoint is $\alpha = 0$ to $\beta = 10$ on $N = 100,000$ vertices with $\epsilon = 0.05$ and $\omega = \log$.

4.2 4.2.3. Performance of the estimator on trees

As mentioned above, from now on we have $\varepsilon = 0.05$ and $\omega(n) = \log(n)$, scaling the $\mathcal{M}_n$ threshold by $\max_{t\in[\varepsilon,1]} D_n(t)$ in all cases. The salient features of the estimator are the following.

4.2 4.2.3.1. Performance vs. the true change point γ

The estimator displays a strong asymmetry with regards to the location of γ, assuming a fixed ε. As evidenced in Figure 4.7, γ close to 0 is easy, and γ close to 1 is hard.

When γ is close to 0, the estimator $\hat{\gamma}$ actually performs quite well. As γ gets closer to 1 however, the estimator is prone to identifying the noise near $t = 0$ as signal of a change point. In other words, as $\gamma \to 1$, $\hat{\gamma}$ displays heavier and heavier bias towards 0. This isn't too surprising. From a mathematical standpoint, the scaling $(1 - t)$ heavily depresses the difference which appears after $t = 0.9$, so it appears small relative to the noise near

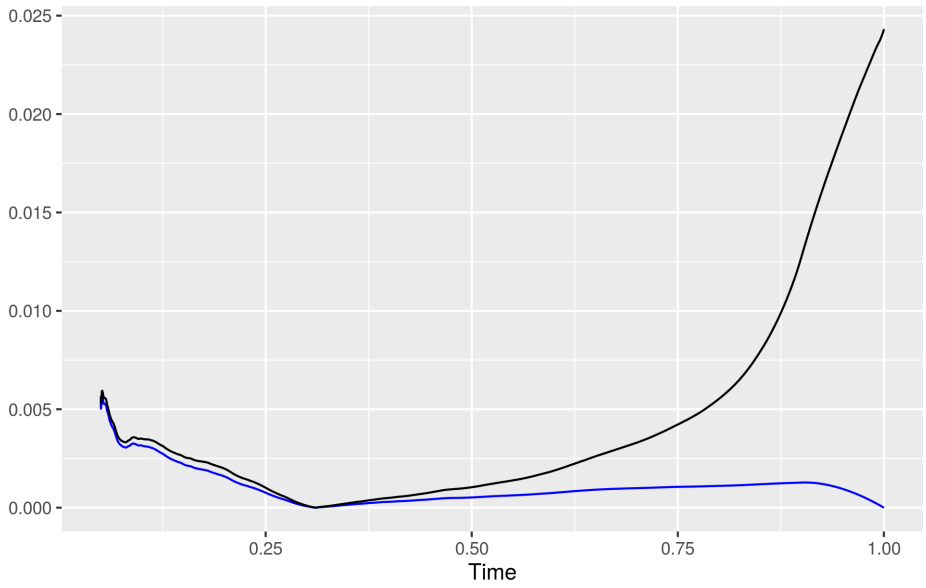


Figure 4.8: Plots of $D_n(t)$ with (blue) and without (black) the scaling $(1-t)$, for $N = 100,000$ vertices and $\gamma = 0.9, \alpha = 0$, and $\beta = 10$.

$t = 0$. From an intuitive standpoint, one would expect that a very late change in network attachment would be very difficult to detect simply because the dynamics are so dependent.

One way to see this is to compare the behavior with that of the unscaled argmax estimator $\tilde{\gamma}$ discussed in subsection 3.3.2, which displays the opposite effect: as $\gamma \to 1$ the estimator biases forwards in time. Without the scaling, $D_n(t)$ overreacts to small changes late in the history of the graph. The situation is made plain by examining scaled vs. unscaled $D_n(t)$ plots for a typical preferential attachment tree with a late change, say at $\gamma = 0.9$, see e.g. Figure 4.8. Despite the deficiency of both approaches, it's still strictly preferable to stick with the scaled estimator. If we knew *a priori* that the change point occurs late, then using the scaled version with large ε, say 0.50, would smooth out the noise appropriately and let $\hat{\gamma}$ pick up the true signal late, see Figure 4.9. This would not be the case with the unscaled estimate—it would still be heavily biased forward, yielding an estimate of very close to 1.0.

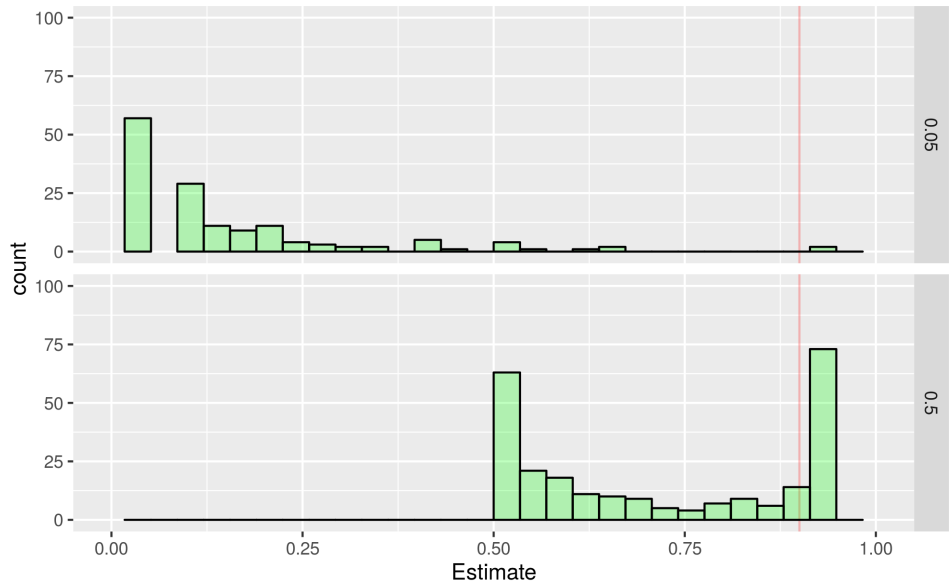


Figure 4.9: Histograms of $\hat{\gamma}$ for $\gamma = 0.9$ with $\epsilon = 0.05$ (top) and $\epsilon = 0.50$ (bottom).

4.2 4.2.3.2. Sensitivity of $\hat{\gamma}$ with regards to $|\alpha - \beta|$

The next natural question with regards to change point detection is how its performance depends on the gap between the pre- and post-change point parameter values. It's not unreasonable to expect that when the change is small, detecting the change is more difficult. Unsurprisingly, this is in fact the case as illustrated in Figure 4.10.

One way to put the relatively difficulty of estimating the time of a small change is to look at the standard deviations for fixed parameter sets. At 100,000 vertices, the sample standard deviation of a sample of 50 estimates of the changepoint from $\alpha = 0$ to $\beta = 10$ is . To achieve the same standard deviation for a change from $\alpha = 4$ to $\beta = 6$, we need roughly vertices.

Of course, this discussion is all modulo other considerations such as the position of γ. Per our discussion in the previous section, it's natural to expect relatively higher precision when γ is small. This is borne out in simulations as well.

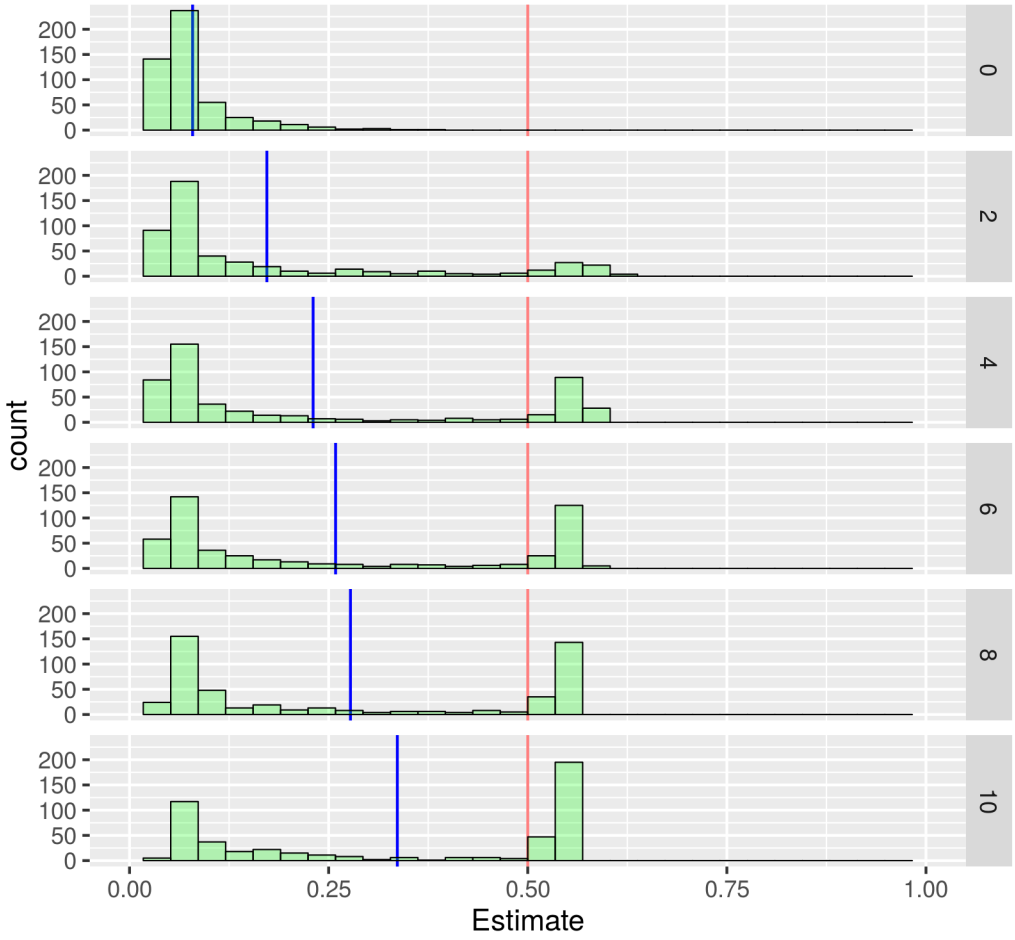


Figure 4.10: Histogram of estimates $\hat{\gamma}$ of $\gamma = 0.5$ with various separations $|\alpha - \beta|$ with $(\alpha + \beta)/2 = 5$ and $N = 100,000$ in all cases. Blue line indicates the mean estimate.

4.2 4.2.4. Extension of $\hat{\gamma}$ to graphs with $m > 1$

A natural extension of our method is to try an adapt it to preferential attachment graphs with multiple edge attachments per vertex, which from now on we shall refer to as preferential attachment *graphs*.

The end goal of this chapter is to see how our ideas extend to real networks, and real networks are rarely trees. Therefore it makes sense to try and establish a baseline for how our model should behave on not-trees by using preferential attachment graphs. A complete analysis of such an extension an entire paper in itself, but let us at least briefly investigate the issue empirically.

There are really two versions of such graphs, distinguished by whether or not they allow for multiple edges. In the large graph limit they are essentially the same (as the probability of a multiple edge $\to 0$ as $n \to \infty$). For our purposes and to match the existing method in the simulation software, we shall assume multiple edges are not allowed. To be precise,

Fix two attachment parameters $\alpha, \beta > 0$, a change point parameter $\gamma \in (0, 1)$, and a system size $n > 1$.

(a) For time $0 < j \leqslant \lfloor n\gamma \rfloor$, a new vertex entering the system at time j connects to m pre-existing vertices *sequentially chosen without replacement* with probability proportional to their current out-degree $+1 + \alpha$.

(b) For time $\lfloor n\gamma \rfloor < j \leqslant n$, the new vertex connects to m pre-existing vertices *sequentially chosen without replacement* with probability proportional to their current out-degree $+1 + \beta$.

4.2 4.2.4.1. The function $D_n^{(k)}(t)$

Obviously, if we are to extract any information from graphs with $m > 1$ then we can no longer rely on counting vertices of degree 1. So let us introduce the notation

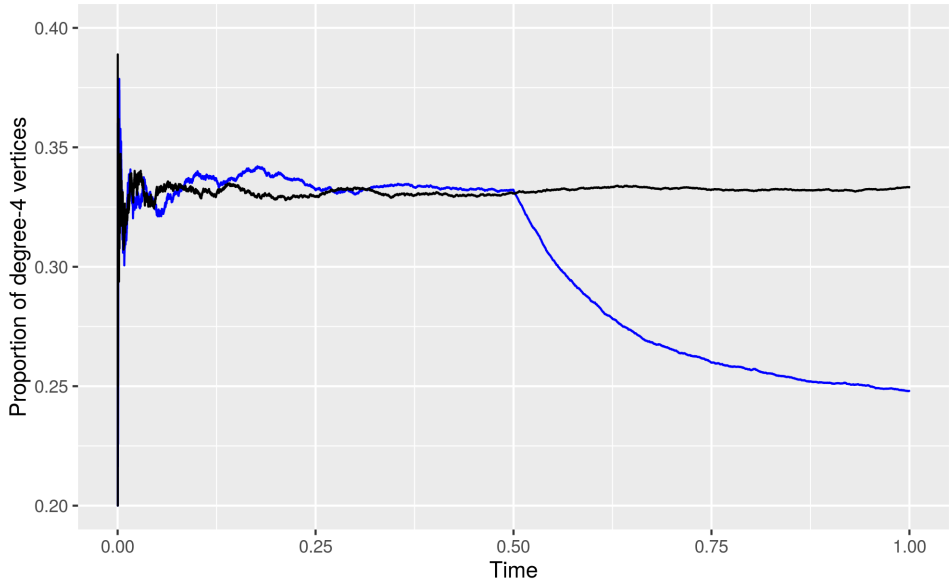


Figure 4.11: The proportion of degree-4 vertices in a preferential attachment graph with $m = 4$ with changepoint at $\gamma = 0.5$ (blue) and without changepoint (black).

$$D_n^{(k)}(t) := |{}_t h^{(n),k} - h_t^{(n),k}|, \qquad t \in [\varepsilon, 1]. \tag{4.2}$$

where ${}_t h^{(n),k}$ and $h_t^{(n),k}$ are the pre- and post-changepoint average proportion of vertices of degree k. It is critical to note that in our definition of $D_n^{(k)}$, we have *not* scaled the absolute difference by $(1 - t)$. This is for a variety of reasons.

First of all, it isn't obvious that $(1 - t)$ is the right scaling for degree-k vertices. The proportion of higher-degree vertices behaves similarly, but slightly differently than in the case of leaves and trees. In graphs with $m > 1$, the proportion of degree-m vertices responds faster to a changepoint, see e.g. Figure 4.11. The exact scaling factor can be calculated by solving a recursion analogous to 3.87, but is annoying. More to the point, even if $(1 - t)$ turns out to be the correct recursion, it would take more work to establish whether or not it's the same for every $k > 1$. Since we will investigate graphs over a wide range of $m > 1$, it's more convenient to omit the scaling.

Secondly, the out-degree of incoming vertices in the real networks we examine in the next section is clearly not constant within a single graph. So for the purposes of establishing

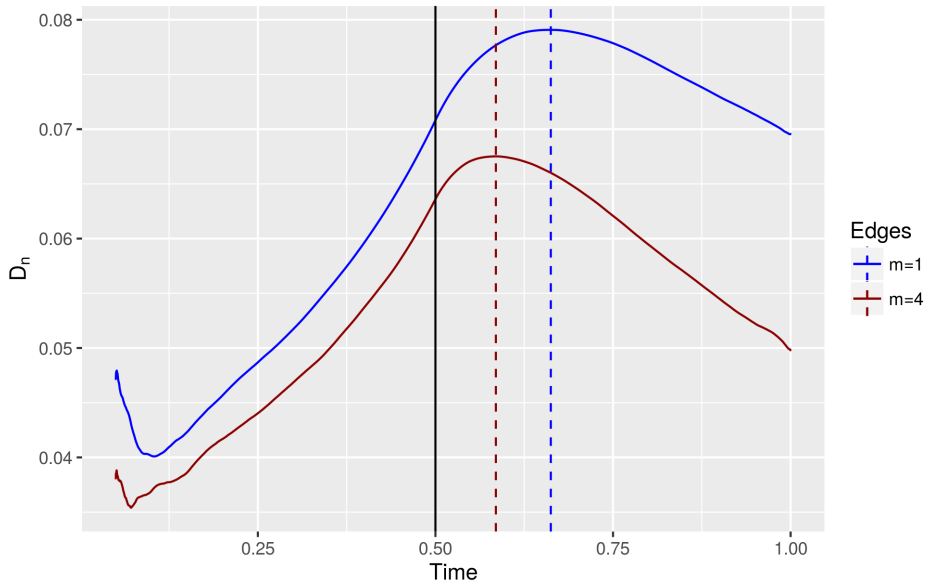


Figure 4.12: Plot of $D_n^{(1)}$ for a preferential attachment tree with $m = 1$ vs. $D_n^{(4)}$ for a graph with $m = 4$ on $N = 100,000$ vertices with change point at $\gamma = 0.5$ from $\alpha = 0$ to $\beta = 10$. Dashed lines indicate argmaxes.

a baseline for real data analysis, it makes sense not to incorporate any scaling even if we had solved for it.

In general though, we would expect $D_n^{(k)}$ on preferential attachment graphs to react to change points much in the same way as $D_n(t)$, the leaf count, does on trees. In fact, it turns out that for a preferential attachment graph with m new edges per incoming vertex, if we set $k = m$ then the function $D_n^{(k)}$ is even better in some ways, e.g. in Figure 4.12.

We saw that with trees the effect of a parameter change is not felt until much later with regards to $D_n(t)$. However, when $m > 1$, each incoming vertex can affect the k-degree vertex count multiple times. Therefore it makes sense that when m is greater, the $D_n^{(m)}$ function not only has a sharper kink where the change occurs, but also that it occurs closer to the true change point rather than after a lag.

Finally, there is the question about what $D_n^{(k)}(t)$ looks like when $k > m$. Essentially, this is the same dynamic as counting degree-k vertices in the preferential attachment tree, where $k > 1$.

The technically-correct answer is that "it depends on the exact parameter values." But in general, we can expect that $D_n^{(k)}$ will have more noise and lag farther behind the true change point. Higher-degree vertices should not feel the effect of a change in the attachment parameter as strongly, simply by virtue of there being fewer high-degree vertices. In addition, the limiting degree distribution arising from linear preferential attachment has finite expectation, so even when weighted by the higher degree, the contribution to the attachment selections by high-degree vertices is necessarily less than the contribution to the attachment selections by low-degree vertices as a whole.

Later when we evaluate real data, we will briefly revisit the hairy problem of figuring out the best k for data which does not have constant attachment out-degree m.

4.3 4.3. Real data

As network analysis becomes more and more popular, the amount of network data available grows exponentially. However, up until now, the common thread running through most network analyses is that they are *static.* Tasks like community detection, clustering, or estimation of various vertex properties generally are performed on snapshots of a graph, instead of over the whole history of the graph.

One side effect of this narrow view is that most available network data either lack time information entirely or contain only coarse records, making it impossible to study the exact attachment dynamics of the model.

We have scraped two citation networks which, to our knowledge, are some of the first large, publicly available temporal networks with fine enough resolution to study on the scale of individual connections. These data are:

1. **The arXiv graph:**

   A citation network of all papers uploaded to arXiv since the first preprint appearing on April 25, 1986. Vertices are papers and two papers have an edge if one of them

cites the other. Citations are linked to arXiv preprints using a proprietary algorithm from Paperscape[2].

2. **The CourtListener graph:**

   A citation network of court opinions from the federal appellate circuit and the Supreme Court. This network was scraped from CourtListener[3] and generously provided to us via Iain Carmichael[4].

Initially, our goal with these data was to directly apply our change point estimator and see what resulted. However, the evolution of these networks is so different from the ideal of preferential attachment that the resulting analysis is mostly pointless. Instead, we aim to conduct an exploratory data analysis of the networks inspired by the approach of our change point estimator.

As a final note, we will generally look at subgraphs of the real networks below, corresponding to certain obvious communities. For arXiv, this means looking at citation networks for papers in the same subject matter category. For the CourtListener data, this means looking at citations of cases within the same court.

Different categories have very different characteristics (e.g. citation conventions amongst biologists might be very different from citation conventions amongst mathematicians). Broadly speaking, we want to isolate changes in a single network's evolution over time due to some structural shift, as opposed to changes due to combining or mixing of phenomena across several different networks. Therefore, in what follows we will generally avoid analysis of the entire network as a whole.

[2] `http://paperscape.org`

[3] `https://www.courtlistener.com/`

[4] `http://github.com/idc9/law-net`

4.3 4.3.1. The raw data

4.3 4.3.1.1. The arXiv graph

This dataset is a citation network from the preprint archive arXiv [5]. As mentioned above, vertices are papers and two vertices are connected if at least one of the corresponding papers cites the other. We analyze the graph as undirected but, intrinsically, it is directed—a paper's outdegree is the number of citations it has to other papers, and a paper's indegree is the number of citations other papers have to it. We will sometimes use this terminology.

The vertices, edges, and basic vertex metadata (arXiv ID, arXiv categories) were provided by the Paperscape data repository[6]. The time data was not included, but was instead scraped additionally using the arXiv API. All code for obtaining the data can be found on the author's Github page[7].

Before we proceed to the exploratory analysis, there is one important note regarding the citation metadata: papers prepared for journal publication generally do not list the arXiv preprint IDs for their references. Instead, they list the citation for the published article. Therefore if one were to try to connect arXiv papers based on their references, one would need to cross-link the journal citation with the arXiv preprint somehow.

Papers on arXiv are grouped into categories which convey which subject matter area the papers concern. The success rate for cross-linking journal citations with arXiv preprints varies widely depending on the category, due mostly to the differing popularity of arXiv across disciplines, and the relative proportion of all published papers which have arXiv preprints.

For example, the field of high-energy physics became active relatively recently, due to technological constraints. Therefore, most papers in that area were published while arXiv

[5] `https://arxiv.org`

[6] `https://github.com/paperscape/paperscape-data`

[7] `https://github.com/yichijin`

Category	Papers	ArXiv refs	Total refs	Ratio
hep-lat	13,243	281,915	344,181	81.91%
hep-ph	91,417	3,099,726	3,836,462	80.80%
hep-th	73,199	2,121,979	2,644,210	80.25%
nucl-th	22,878	528,484	834,066	63.36%
astro-ph	179,168	3,205,223	8,286,626	38.68%
cond-mat	181,732	1,530,928	5,620,147	27.24%
physics	68,050	237,050	1,641,857	14.44%
math	213,993	356,700	4,388,131	8.13%
cs	83,712	50,714	1,568,814	3.23%

Table 4.1: Paperscape scraping success rates for selected categories.

was online and consequently most high-energy physics papers have arXiv preprints[8]. Mathematics, on the other hand, has a large corpus of still-influential papers from the early 20th century for which no preprints exist. Therefore, for a typical math paper, a smaller proportion of references can be successfully linked to an arXiv preprint.

We refer the reader to the Paperscape and Github links above for more details regarding the data scraping methodology.

4.3 4.3.1.2. The CourtListener graph

CourtListener is an online, open-source project which archives and organizes court opinions from various courts around the country. Our data set covers the federal appellate circuit and the supreme court only. As in the arXiv data, we will analyze the graph as undirected but it is intrinsically directed.

The vertices in the CourtListener graph are court opinions running all the way back to February 14, 1792. Two vertices have an edge between them if one of the court opinions cites the other. We treat the graph as undirected. Vertices also come with a couple points of metadata, most importantly the (unique) CourtListener ID, the publication time, and the court ID where the opinion was written.

[8]Also, arXiv was developed at Los Alamos National Laboratory by physicists.

Similar to the case of arXiv, CourtListener has a custom codebase for automatically detecting citations to court opinions and linking them to documents in their database. However the CourtListener data is slightly messier than the arXiv data, for two reasons:

1. CourtListener does not detect citations to or archive statutes, law review articles, or other resources.

2. CourtListener does not currently publish (or possibly know) the percent coverage (in their database) of all opinions ever written. This is partially due to the fact that CourtListener does not scrape all of the citation data themselves—rather, they rely somewhat on donations from partner sites. Therefore the scraping success rate is subject to inconsistencies due to the different algorithms employed across the partner sites.

For our purposes, we will generally treat the data sets as complete, 100% samples of court opinions written in each court. However, in later analysis we are careful to keep the fact that we don't know the exact recovery rate in the back of our minds.

We refer the reader to the CourtListener website[9] for more details about the raw data.

4.3 4.3.2. Does it look like preferential attachment?

As alluded to in the introduction, the main takeaway from this section is that these real networks look close enough to preferential attachment in the $t = 1$, aggregate view. But when we inspect their history, they clearly don't. Let us dive straight into it.

4.3 4.3.2.1. The arXiv graph

One of the motivations for development of preferential attachment is that it was one of the first temporal models capable of producing a power-law degree distribution.

[9] `https://www.courtlistener.com/about/`

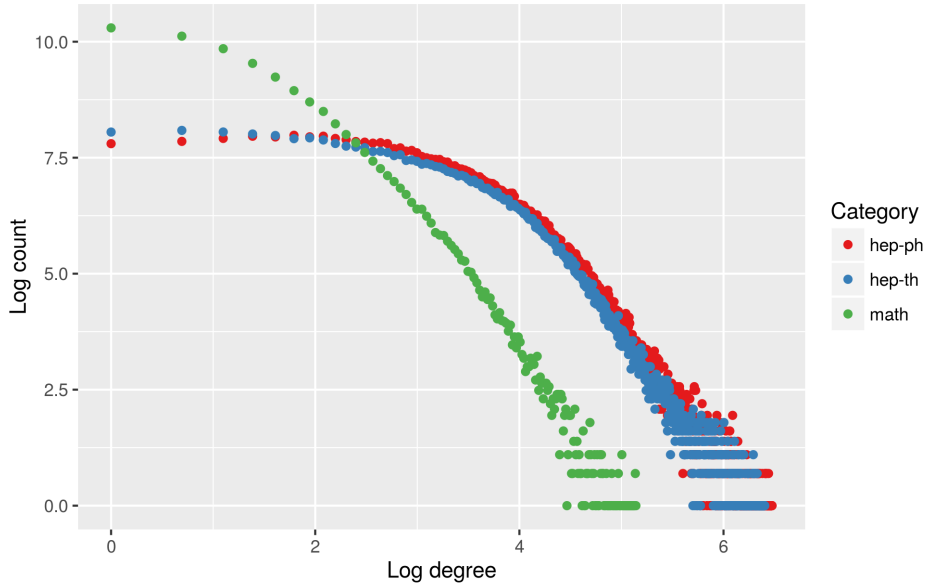


Figure 4.13: Log-log degree distribution for selected arXiv categories.

The degree distribution on the arXiv graph really depends on what category we are looking at. No, the distribution is not strictly a power law for the entire range of degrees. However, in most cases we see a strong linear trend on the right side of the log-log plot indicating a power law tail. A lot of this can be explained by the fact that in these real networks, there is additional randomness arising from the random number of initial out-degree of vertices corresponding to how many citations to existing preprints a new arXiv preprint has. Because this initial out-degree rarely exceeds say, 200, this randomness only affects the lower portion of the degree distribution so its effect is not felt in the tail. This is consistent with what we see in most citation networks.

Also, the aggregate statistics are generally in line with a comparable simulated preferential attachment graph, see Table 4.2.

In particular, the order of magnitude of the maximum degree and the clustering coefficient are roughly similar. If these were the only features we looked at, we could say with reasonable confidence that these graphs arose out of a preferential attachment-type of dynamic. Another takeaway from this table is that there are clear dominating categories on the arXiv graph. Having been created by physicists, arXiv sees the heaviest use from that

Category	Average degree	CC	Edges	Maximum degree	N
astro-ph	32.17	0.01	2,972,371	8,060	184,815
cond-mat	15.54	0.01	1,298,041	3,185	167,042
hep-ph	39.58	0.02	2,315,801	3,005	117,028
hep-th	37.05	0.02	1,917,902	9,192	103,542
math	5.93	0.02	421,043	684	142,117
PA, $m = 8$	16.0	0.00	799,964	1,332	100,000
PA, $m = 1$	2	0	99,999	180	100,000

Table 4.2: Aggregate statistics for the arXiv graph (selected categories).

field. Indeed, we will focus on a select few from now on, since those are the ones most likely to adhere to a uniform attachment dynamic over all of its vertices.

The most defining characteristic of pure preferential attachment is that incoming vertices attach to existing ones with probability proportional to the degree of the existing vertex. But before discussing even that, the most obvious way in which the these arXiv communities diverge from theory has to do with the initial out-degree—which in citation networks corresponds to the number of references to other (existing) papers within each paper. The fact is that not all vertices join the graph with the same out-degree. And not only that, but the *average* initial out-degree increases with time, see e.g. Figure 4.14.

It's clear why this is. Even though there is no reason to expect the average number of total references in each paper to be increasing over time, recall that the arXiv citation network tracks citations from arXiv papers to *other arXiv papers*.

What we have is sample selection bias in the early days of arXiv. arXiv was officially launched in the summer of 1991. Initially, there were only a very small corpus of papers on arXiv. This steadily grew as time went on. As arXiv matures further and collects more preprints of cutting-edge research, more and more arXiv preprints will contain citations to those papers already on arXiv. This narrative is borne out simply by looking at the density of papers over time. See for example Figure 4.17.

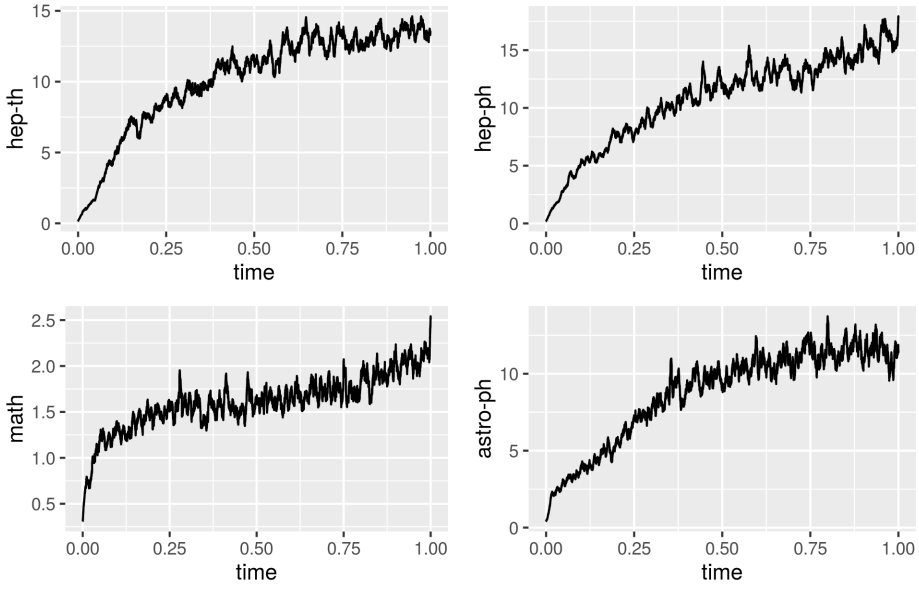


Figure 4.14: Time series of initial out-degree of per paper for selected arXiv categories, smoothed by moving average over 2000 papers.

One can imagine that there will come a day when all relevant past research have arXiv preprints so the trend in this series will flatten out, but the continuing upward trend at this point in time tells us that this process is still under way.

4.3 4.3.2.2. The CourtListener graph

The CourtListener graph shares many of the same similarities and departures from preferential attachment as the arXiv data. And similar to as in the arXiv analysis, we will generally look at subgraphs corresponding to natural communities in the data, which are the appellate courts.

First off, most of the appellate court graphs have an approximate power law degree distribution. It's not as strong as the same effect in arXiv citations, but one can imagine several explanations for why court citations may have a lower ceiling with respect to very-high-citation-count court opinions.

Furthermore, in this picture we can see some evidence of a slight kink in the degree distribution—this is a common feature of many power law graphs including social networks.

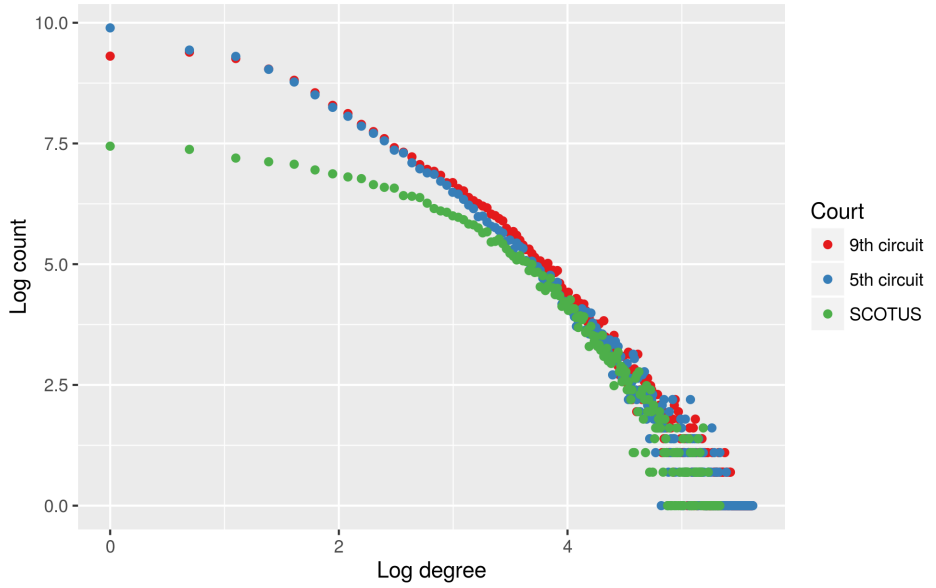


Figure 4.15: Log-log degree distribution for selected courts.

In words, the degree distribution scales a certain way from 0 up to a certain degree k, and then abruptly switches over to a different scaling for degrees $k+1$ and beyond.

There are a couple ways to explain this. A kinked distribution is especially common when viewing directed networks as undirected, where the out-degree has one scale and the in-degree has another. For example, in scientific preprints a paper will rarely ever cite more than 200 papers, but it is relatively common for a paper to have over 200 citations *to it*. When these degree counts are combined and plotted, this manifests as a kink where one scale stops and the other begins to dominate.

We also saw in the previous section that the features of the arXiv network depended heavily on the subject category in question, but also that a lot of the arXiv network evolution was driven by arXiv's relative newness. There's reason to suspect that things might be different in the CourtListener data.

For one, there is not as much of a reason to suspect that different appellate courts have widely-differing citation habits. After all, all courts handle all sorts of cases, so there isn't a differentiation of training across courts like there is across physics versus computer science, for example. This is evident in the aggregate statistics across districts, see Table 4.3

Court	Average degree	CC	Edges	Maximum degree	N
ca1	13.09	0.02	158,911	1,086	24,279
ca2	11.06	0.01	219,269	1,078	39,651
ca3	10.05	0.01	187,838	677	37,373
ca4	6.23	0.00	188,951	5164	60,637
ca5	9.21	0.00	442,008	3197	95,975
ca6	10.29	0.01	267,130	661	51,905
ca7	13.89	0.01	330,778	837	47,638
ca8	10.08	0.01	286,890	725	56,898
ca9	10.66	0.01	480,777	1131	90,211
ca10	10.50	0.01	242,325	1209	46,141
ca11	10.92	0.00	207,832	3976	38,074
scotus	18.89	0.02	234,155	322	24,795
PA, $m = 5$	10	0.00	249,985	891	50,000
PA, $m = 1$	2	0.00	49,999	451	50,000

Table 4.3: Aggregate statistics for the CourtListener graph.

Secondly, CourtListener has worked diligently to digitize as many court opinions as possible, even those going back close to the inception of the circuit court system. Plus, most of the circuit courts have been around for close to 150 years—an eternity when compared to arXiv. Therefore we would expect to see less of an artifact stemming from early sample selection bias, which should translate to a more stable picture with regards to new-vertex out-degree.

The evidence in Figure 4.16 may or may not support this hypothesis. On one hand, there do appear to some circuits which have relatively constant initial out-degrees. For example, the 7th Circuit Court of Appeals displays a lot of variation over the past century but doesn't display a clear upward trend. On the other hand, some of the courts have very different patterns.

The Supreme Court has a very strong upward trend to its initial citation counts, and the 9th Circuit Court of Appeals has a disconcerting jump in average citation count roughly between the years of 2000 and 2010 which is not shared by any other court system.

It's not clear if these counts are due to a real structural change within the court system, or whether they are artifacts of how the data were collected. At any rate, it's safe to say

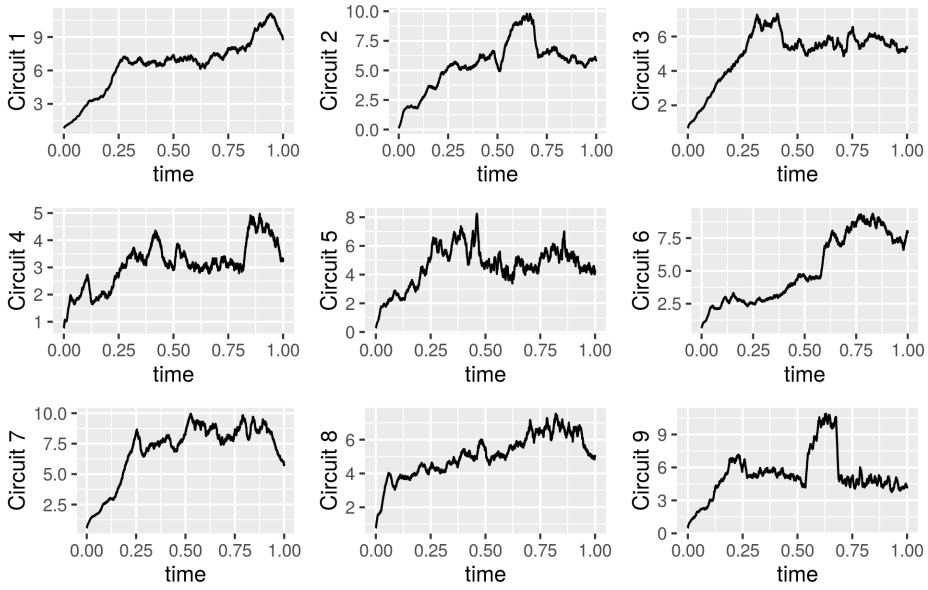


Figure 4.16: Time series of initial out-degree of per paper for selected appellate courts, smoothed by moving average over 1000 cases.

that even though these court opinion graphs may be the closest real data we have to pure preferential attachment, they remain a far cry from it.

4.3 4.3.3. Analysis of the network history

Let's take a look at some of our graphs and see what we can discover from their histories.

Technically, we *can* run the $\hat{\gamma}$ change point estimator on these data. But it's clear that doing this makes little sense. As noted above the evolution of these networks does not look like preferential attachment, and furthermore there are various other issues with trying to adapt $D_n(t)$ to these graphs.

Taking a step back though, we should be able to apply our general strategy to other functionals of the graph history. The general approach is simple:

1. We identify a function of the graph history which *in theory*:

   (a) is constant in the absence of a change point

   (b) is not constant in the presence of a change point

2. We calculate the empirical version of that function and see whether it agrees with (a) or (b) and conclude.

So essentially, as long as we can imagine a function of the graph history which ought to be roughly constant over the history of the graph in the absence of a major structural change, we should be able to perform some type of change point detection using the empirical version of that function.

Of course, this is all in theory. The reality is that real world graphs can break from this ideal in many ways. For example, external shocks to a network may only affect a small fraction of the vertices. Or probably even more likely, real structural changes to a network might occur slowly, even incrementally over a non-trivial span of the network's lifetime.

What we will see in the analysis of the arXiv and CourtListener data is that all of these are possibilities, and abrupt changes are hard to come by. We'll focus on three functions of the graph history:

1. The k-degree vertex proportion
2. The initial out-degree sequence
3. The average degree of the initial neighbors of incoming vertices

There are many possible extensions of these statistics. A discussion of those is out of the scope of this thesis. However, we hope that this will stand as a proof-of-concept for the potential of this way of observing network change.

4.3 4.3.3.1. The time scale of real life

In all the plots that have been presented thus far, the time scale on the x-axis is not the time scale of real life. Why? Because the time index $t \in [0, 1]$ is based on vertices' relative *order of appearance* in the graph. Whether or not the second paper in arXiv appear 1 month or 1 year after the first paper makes no difference to the fact that it was the *second* paper.

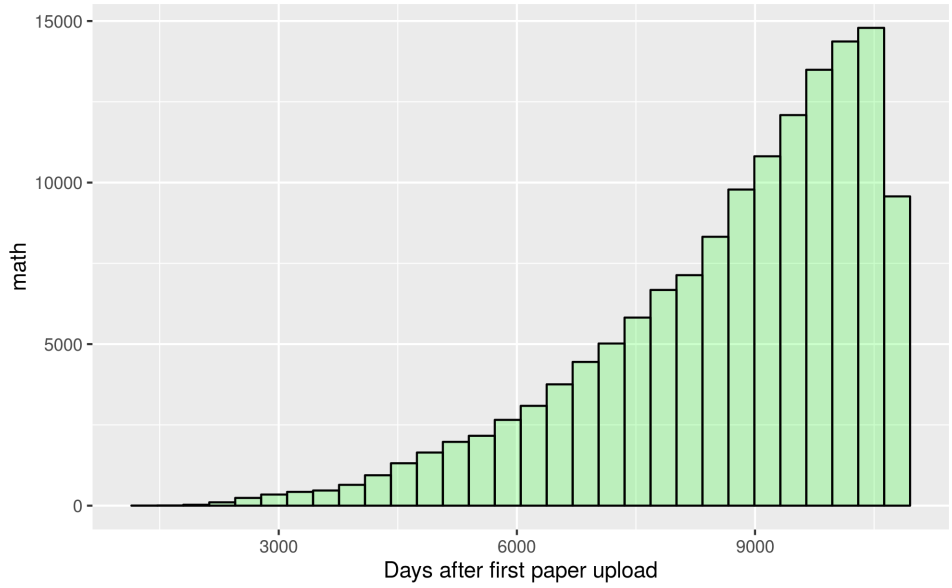


Figure 4.17: Distribution of paper appearance times in arXiv categories *hep-ph* versus *math*.

Therefore for the purposes of identifying structural breaks in the graph with real life events, we must re-scale all our plots to take the *time* of appearance into account. This time scale is different for each subgraph we have investigated, reflecting the relative "popularity" of each arXiv category or appellate court over the years.

We will essentially stretch each series we've presented thus far by this time scale. For the sake of readability, we'll also normalize time to *days since the appearance of the first vertex in the overall graph*. The difference is generally not too large in most graphs, but will allow us to easily interpret features in the series. See Figure 4.18 for a case where the distortion is relatively high—in the `math` arXiv category as shown in Figure 4.17.

4.3 4.3.3.2. The large hadron collider?

The arXiv category `hep-ph` stands for "High-energy physics - Phenomenology." This is the application of theoretical physics to high-energy experiments. This is contrasted with the `hep-th` category which deals mostly with the development of high-energy physics theory.

The initial out-degree plot for `hep-ph` is quite plain at first glance, see Figure 4.19

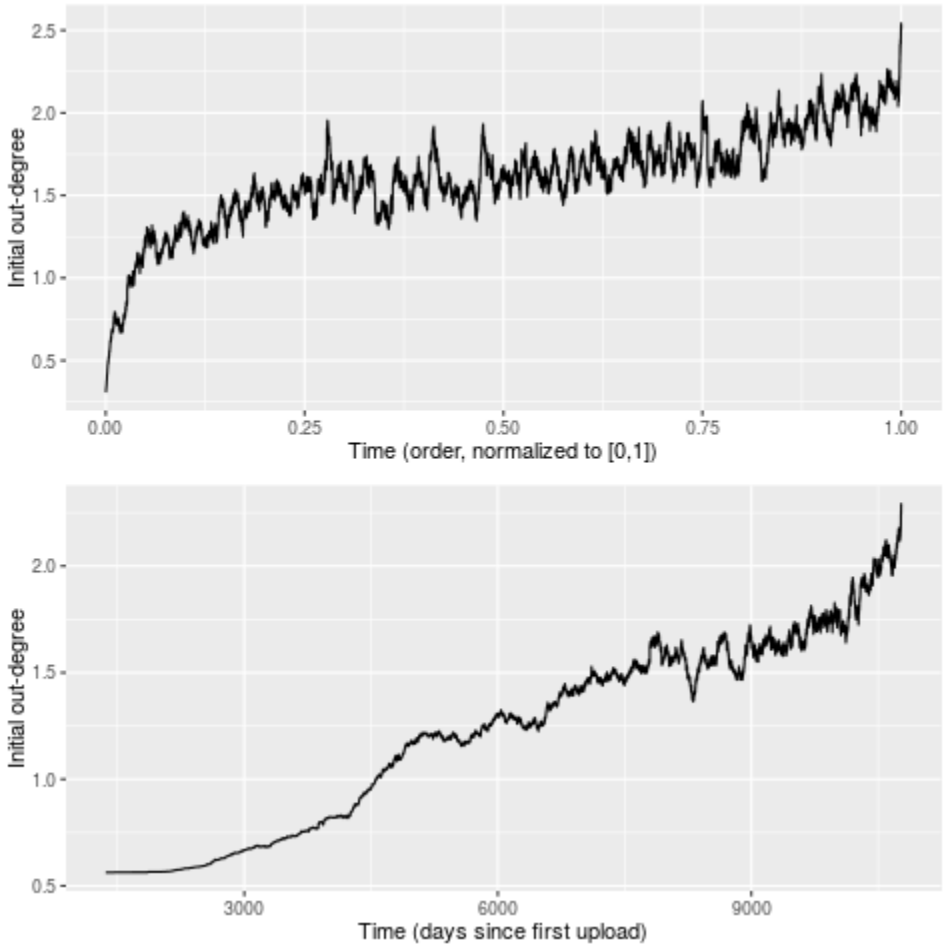


Figure 4.18: comparison of the initial out-degree series for arXiv category *math* plotted on the order-based time scale (top) versus the real-life time scale (bottom).

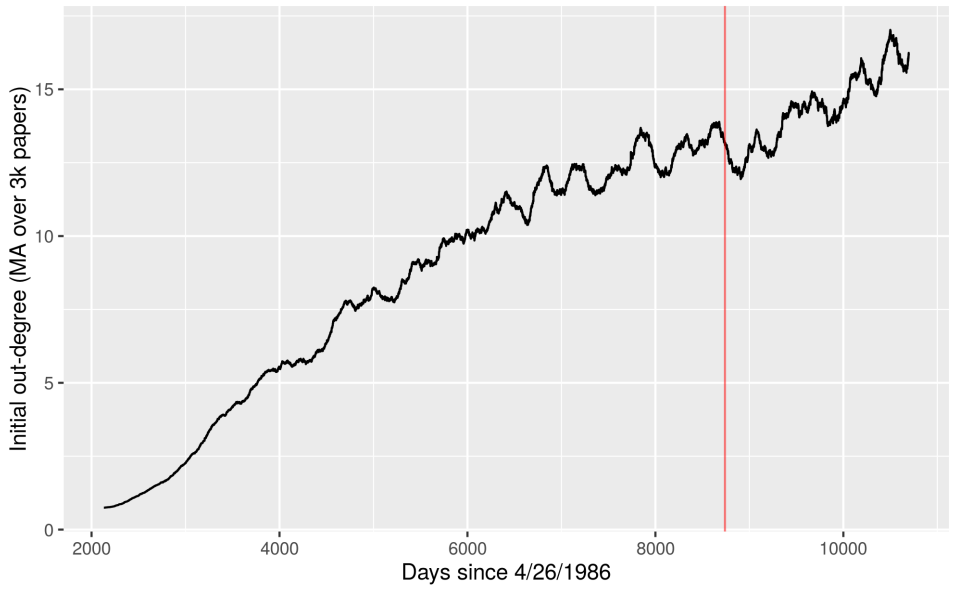


Figure 4.19: Initial out-degree for the *hep-ph* category, moving average over 3000 pages. Red line indicating March 30, 2010.

But, referring back to the histogram Figure 4.17, we see an uptick in papers starting around 9000 days, or roughly Winter 2010.

It turns out that this coincides with the first operational run of the Large Hadron Collider (LHC), the world's largest and most powerful particle collider, which had its first research run from March 30, 2010 through February 13, 2013. After February of 2013 the collider shut down for a period of roughly 2 years to undergo planned upgrades, after which it restarted for a second run.

A priori, we can imagine that this would have a significant effect on the `hep-ph` citation network. After all, the main source of data for this field comes from particle colliders like the LHC.

However, this isn't really the case in our data. If we squint, we might be able to see some slight evidence of this in the initial out-degree plot Figure 4.19, manifesting in a small drop in the average initial out-degree. This would imply that as the LHC came online, papers began citing fewer sources on average. This might be the case if the LHC became the sole

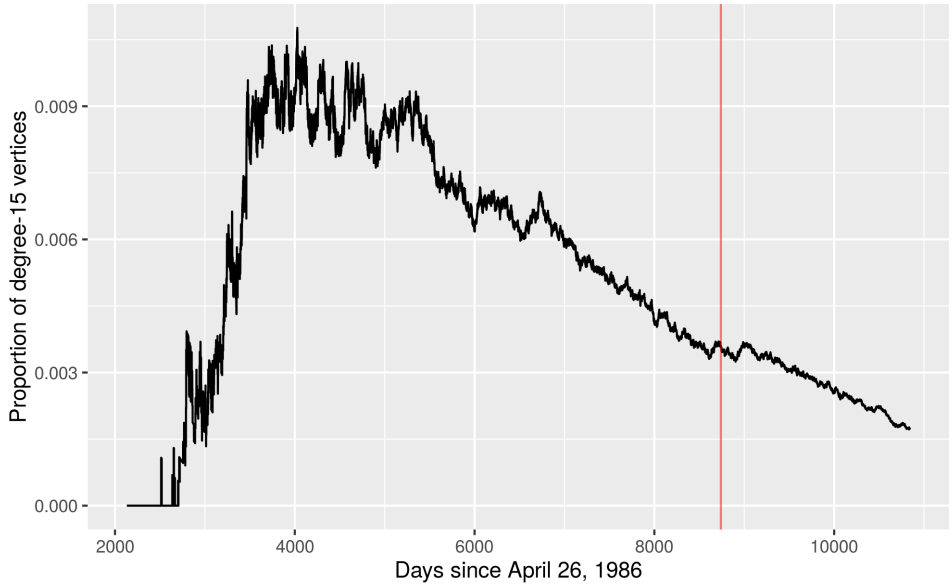


Figure 4.20: Proportion of degree-15 vertices in the *hep-ph* category. Red line indicating March 30, 2010.

focus on research for a couple years and a few landmark papers drove all new research. This may or may not be the case.

This narrative is however also (weakly) suggested by the plot of proportion of degree-15 vertices in the graph, Figure 4.20. The downward trend in the majority of the plot is consistent with a growing graph and steadily increasing average degree over time, as just discussed. However we do notice a temporary pause in the downward trend around the time of the LHC's construction. One possible explanation for this is that new research around this time was concentrated on citing a few high-impact research papers, or simply citing fewer due to availability of results purely from analysis of novel data from the LHC.

Finally, one would hope that this narrative would be backed up by evidence from examining the *degree* of the attached-to vertices for each new arXiv preprint. For instance, if all research is concentrated in a couple high-impact papers, then those papers should be oft-cited and also sport a very high degree. So if we plotted the average degree of the 3 highest-degree papers cited by every paper, we would hope to see an uptick during periods of intensive research on a single topic.

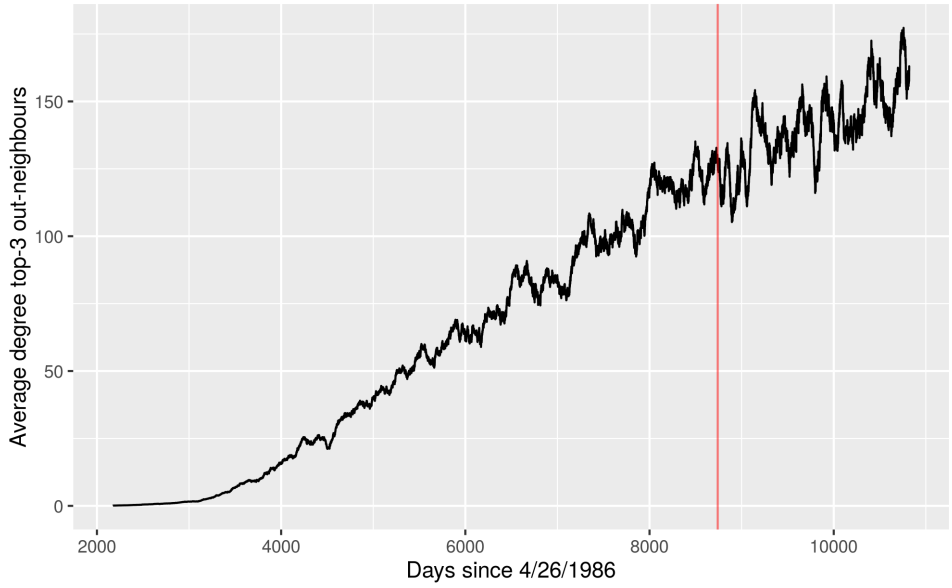


Figure 4.21: Average degree of top-3 out-neighbors, moving average over 1000 vertices. Red line indicating March 30, 2010.

From Figure 4.21 though, this isn't obvious at all.

The main takeaway from this analysis is that these graph series are suggestive and do tell us a good amount of information about the overall evolution of the graphs, but pinpointing abrupt change points is difficult. A change point like the LHC coming online is about as abrupt of an organic change as one can expect, yet even that may affect only a small fraction of papers and occur more gradually than we would like for change point detection.

Other arXiv categories show relatively normal and stable evolution, so let us move on to a brief discussion about the CourtListener data.

4.3 4.3.3.3. The 4th and 9th circuit courts

As mentioned earlier, the CourtListener data is messy. There are bound to be some artifacts due to data collection which nevertheless present as structural breaks. And we see those right off the bat. For these court data, we've normalized time zero to be August 4, 1924, which is the date of appearance of the first case in either the 4th or 9th Circuit Courts

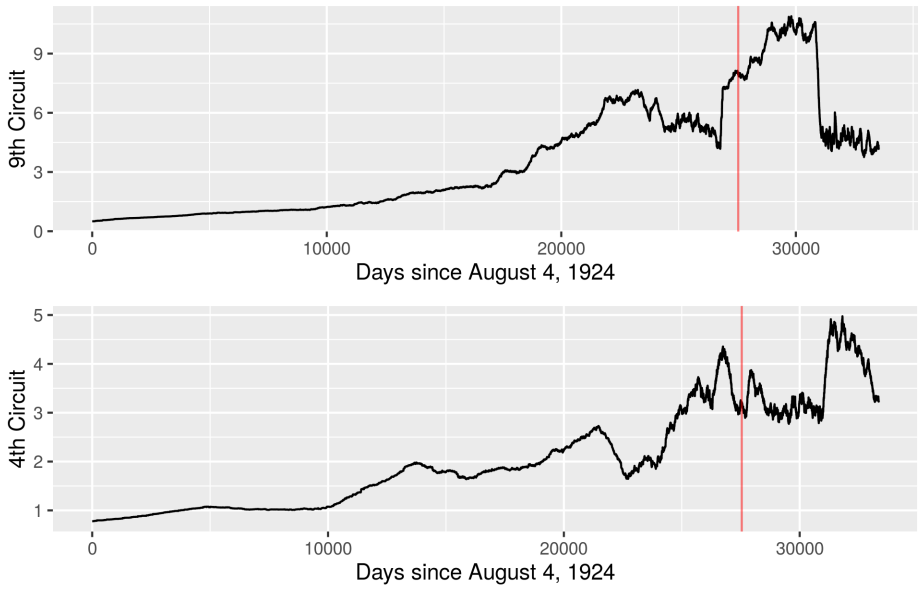


Figure 4.22: Initial out-degree of court case citations from the 4th and 9th Circuit Court of Appeals, smoothed over 1000 cases. Red line at January 1, 2000 for reference.

(including the Supreme Court cases would shift time zero to February 14, 1792, which would make it difficult to see the action in modern times).

Since this plot are smoothed over 1000 cases, the abrupt jumps occurring near 2000 are very surprising. Politically and legally, there are no obvious reasons why we should expect such shocks. In addition, these dramatic changes do not occur in the other Circuit courts.

No matter what the reason for these shocks though, let us take a look and see if they are reflected in our other statistics.

In the citation appearance histograms (Figure 4.23), we see some curious patterns which aren't fully consistent with our change point from the previous chart. What's clear is that there is very wide variation in the number of citations appearing in the data set from year to year. Again, there is not obvious political or legal reason to expect these jumps.

We see that in these breaks are strongly reflected in the proportion of degree-6 vertices in the graph. In the 9th Circuit for example, we see a drop in the proportion of degree-6 vertices just as the average initial out-degree increases.

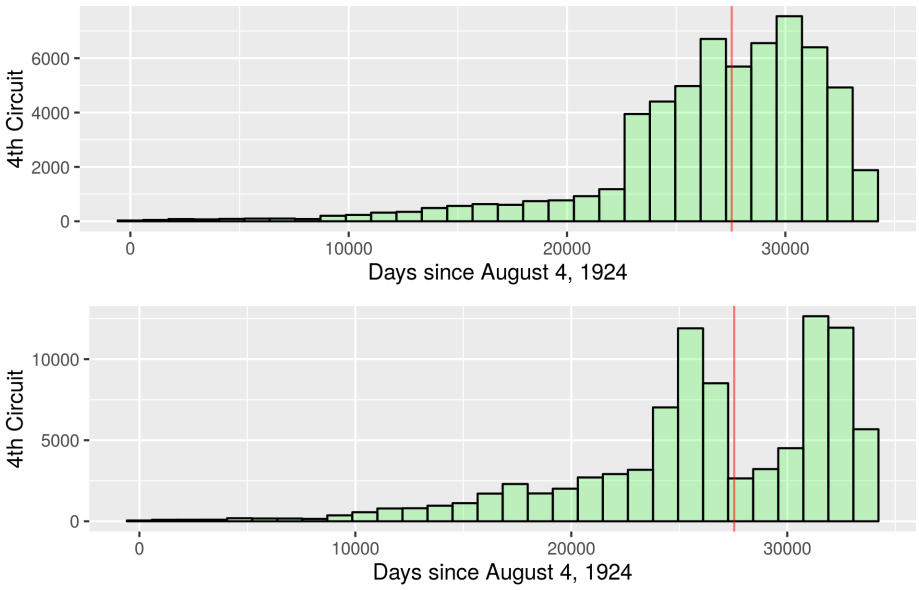


Figure 4.23: Distribution of citation appearance times. Red lines at January 1, 2000 for reference.

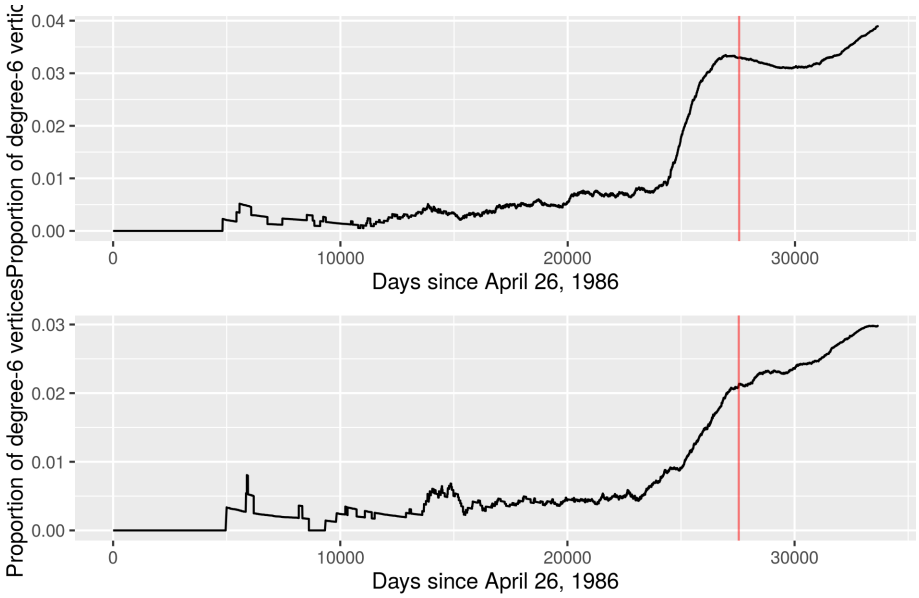


Figure 4.24: Proportion of degree-6 vertices. Red lines at January 1, 2000 for reference.

All in all, this suggests that these simple graph functions move in concert to convey roughly the same story about the graph. Unfortunately, we don't have enough background detail about the CourtListener data to inform the reader about the root causes of this

4.4 4.4. A note about code

It goes without saying that there was an extensive amount of code written to perform these analyses. The interested reader can find all relevant code on the author's github page[10].

Python was used to perform all the heavy lifting in this section, leaning especially on the `NumPy` and `graph-tool` packages. `BeautifulSoup` and `requests` were used to scrape data from arXiv and CourtListener.

All graph calculations were performed using the `graph-tool`[11] suite of graph analysis tools alongside the numerical Python library `NumPy`. For temporal networks, `graph-tool` is especially useful due to its ability to "view" subgraphs in a $O(1)$ operation without occupying extra memory. We refer the reader to the documentation in the site above for more information.

4.5 4.5. Summary

The main takeaway from these analyses is that looking at functions of the graph history is a must when setting out to explore temporal graph data.

In the case of plain vanilla preferential attachment, we show that the estimator works well and has simple and promising extensions to preferential attachment with multiple edges. It's not a stretch then to believe that it would extend naturally to multiple change points as well.

Unfortunately, real world data is not generally so well-behaved. Nevertheless, we showed that looking at very simple graph functions can reveal a great deal of understanding about

[10] `http://www.github.com/yichijin/pa-changepoint`

[11] `https://graph-tool.skewed.de/`

what's going on structurally behind the scenes. At this point the approach raises more questions than answers, but we feel that it will be possible in the near future to extend our change point methodology to adapt to these messier situations.

CHAPTER 5

Decreasing cascades on scale-free graphs

5.1 5.1. Introduction

This chapter seeks a simple answer to a complicated question: how does information spread on a social network?

The dynamic we would like to study in this chapter is the propagation of information through a scale-free network—specifically, a very specific type of propagation inspired by "retweet" dynamics on the social networking platform Twitter. In contrast to our change point work, we are less interested in the growth dynamic of the underlying network and more interested in the cascade on top of the network.

Almost all person-to-person interaction on the internet is, by definition, carried through on social networks. One way that interactions on the internet differ from interactions in reality is that the internet facilitates large-scale, instantaneous propagation. In essence, social networks make each user their own personal media outlet.

On most social networks such as Facebook or Twitter, interactions broadly fall into one of two classes which we will call *engagements* or *broadcasts*. Engagements are selective interactions between two users, such as private messages on Facebook or direct messages on Twitter. Broadcasts are exactly what the name implies–indiscriminate blasts to all of a user's contacts on the network.

We are interested in studying the dynamics of how information disseminates on a social network through broadcasts. When a single user authors content on a social network and broadcasts it to a neighbor, that neighbor can either choose to ignore it or to re-broadcast

it to their neighbors. This propagation history traces a subgraph on the originating user's social network.

Cascades have been extensively studied, and researchers have proposed countless plausible mechanisms for generating them, see Chapter 2. However, in light of new results by [52] on the *shape* of viral cascades, we believe a new, simple model of information cascades is of relevance.

In thinking about models for cascades on social networks, a few empirical observations must be taken into account. The first is the intuitively obvious fact that the vast majority of them are tiny. In layman's terms, most content on the internet is not extensively shared. However, a very small fraction of cascades break the mold and propagate explosively across the internet, or *go viral*. Any cascade model must be flexible enough to generate cascades at either extreme.

The second observation is that is that in very large (viral) cascades, the *shape* of cascades does not match what is predicted by simple epidemic models (see Section 2.3.2). To recap, classical models for cascades all predict that large cascades will generally fall into one of two extremes: those that are truly viral, reaching a long distance away from its source and many users at each distance; and those that are simply broadcasts, reaching only users within one or two hops from the source. In [52] this is summarized using the following concept.

For each cascade T we can associate a measure of its "viralness" by the average shortest path distance between nodes:

$$\nu(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$$

where i, j index the nodes of T and d_{ij} is the distance from node i to node j.

Most simple cascade models predict that either $\nu(T)$ is close to 2 or very large (depending on the size of the graph). In terms of familiar constructs, if a branching process is supercritical, then if it survives past the first generation it will tend to survive for many

generations and have large $\nu(T)$. If on the other hand a branching process is subcritical or supercritical conditioned on extinction, it is unlikely to survive past a handful of generations and thus have $\nu(T)$ close to the minimum value, 2.

In practice, [52] notes that actual social network cascades exhibit a wide range of viralities rather than being bimodally distributed at the extremes, and this is backed up by many other studies. Therefore our challenge is twofold. First of all, how can we capture this range of behaviors? Second of all, given the simple way retweeting works in reality, how can we accomplish our first goal using the simplest possible model?

A complete answer to this question is out of the scope of this thesis, but we endeavor to take the first step by proposing a simple branching process model which, we argue, is a reasonable candidate for generating these types of flows.

5.2 5.2. A cascade by a branching process

Our proposed model is an easy modification of conventional Galton-Watson processes which nevertheless produces fairly realistic cascades.

Recall that in Section 2.3.2 we explained how modelling a retweet cascade using a GWBP can only produce large cascades of a very specific size because the offspring distribution must be constant across all individuals and generations. But there's a simple observation which suggests an easy fix.

From a tweet's perspective, each tweet actually faces a different graph upon which to propagate when it comes time to be retweeted. Each tweet author has a different follower graph which might even change over time. Some authors have millions of followers, while some may only have thousands or hundreds. It makes no sense to enforce a constant offspring distribution across individuals. But more than that, there is an implicit decaying of relevance (and therefore probability of retweeting) as the cascade gets farther from the root. Thus suggests even more reason to relax the constant offspring distribution assumption and take us into the world of inhomogeneous branching process.

To be fair there are a lot of fancier branching process models out there with countless parameters and rules designed to replicate real-world phenomena. But as mentioned above, one of our goals is to show that, when correctly tuned, a plain vanilla branching process works surprisingly well in generating these cascades. First we describe the process on a graph and then discuss the implicit assumptions. It will turn out that this process is essentially equivalent to a simple variant of a supercritical Galton-Watson process.

5.2 5.2.1. Decreasing cascades

Suppose we have an arbitrary scale-free graph G with vertex set V. This cascade explores G in discrete time through a set of active nodes, tracing a tree structure in the following way.

At time n, active nodes infect their unexplored neighbors independently with probability p_n. Once a node is finished infecting its neighbors *or has failed to become infected*, it cannot infect any more nodes. The set of nodes which were successfully infected during time n then becomes the set of active nodes of time $n+1$. In other words, if a person is susceptible but fails to become infected, then we consider them immune from all further infection attempts. Thus the infection path is always a tree.

More formally, at each time n, V will be partitioned into sets A_n, E_n, U_n where A_n is the set of *active* vertices in the graph, E_n is the set of *explored* vertices, and $U_n = V - (A_n \cup E_n)$ is the set of *unexplored* vertices.

1. To initiate the cascade at node v_0, set $A_0 = v_0$, $E_0 = \emptyset$ and $U_0 = G - \{v_0\}$ and let $\{p_n\}_{n \geqslant 0}$ be a sequence of probabilities with $p_n \downarrow 0$.

2. For each n, initiate $A_{n+1} = E_{n+1} = \emptyset$ and iterate over $v \in A_n$ to populate A_{n+1} sequentially as follows.

   (a) Try to infect each $w \in N(v) - E_n$ independently with probability p_n.[1]

   (b) If w is successfully infected, then $A_{n+1} \to A_{n+1} \cup \{w\}$

[1]Throughout, we use the notation $N(v) =$ the neighbors of vertex v and $N(A) = \{\cup_w N(w) : w \in A\}$.

(c) If w fails to become infected, do nothing.

3. After all $v \in A_n$ have been explored, then move all $w \in N(A_n) - E_n - A_{n+1}$ to E_{n+1}.

4. Move $E_{n+1} \to E_n \cup A_n$

Note that if an unexplored node $v \in E_n$ has k active neighbors, then it becomes infected with probability $1 - (1 - p_n)^k$.

To translate into Twitter terms,

1. A_n is the set of users who have seen the retweeted content and are currently deciding whether or not to retweet it.

2. E_n is the set of users who have seen the retweeted content and decided not to pass it on.

3. U_n is the set of users who have not yet seen the retweeted content.

4. p_n is the *transmission probability*—the probability that, at distance n away from the source (or, equivalent, after n time units have passed), an user will retweet to his/her followers.

The implicit assumptions in this model are the following:

1. **Transmission probability decreases with the distance from the root:**

   This belief actually reflects another implicit assumption on the underlying graph on which the cascade spreads: users closest to the source node have the highest interest in the content produced by the source, and therefore the highest probability of passing it on.

2. **Transmission probability is constant over all individuals at a given distance from the root:**

   In reality, this assumption is almost certainly false. However, it isn't unreasonable to imagine that this is a good approximation.

3. **The discrete-time view is enough:**

   The motivation for devising continuous-time cascade models is the belief that viral cascades' transmission dynamics depend on the time elapsed since the content was produced. Although our model has no time component, the decreasing probability scheme captures to some degree the belief that the greater the age of a tweet (and therefore the farther it spreads), the more its attractiveness decays.

Empirical verification is difficult for some of these assumptions due to technical reasons (see [104] for an example), but nevertheless we believe that the model simplification it allows is significant enough that it is warranted. Coupling to a branching process framework will allow us to show that this cascade has several nice properties.

5.2 5.2.2. The branching processes approximation on a graph

Due to the tree structure of the cascade, it isn't hard to see that the cascade resembles a sort of branching process with a special sequence of offspring distributions. The gist of the resemblance is as follows: starting with such a graph G, an arbitrary source vertex v_0 on that graph, and a decreasing probability sequence $\{p_n\}_{n\geqslant 0}$, we infect a Binomial$(N(v_0), p_0)$ number of neighbors, where $N(v_0)$ is the total number of neighbors of v_0. Then, a newly-infected vertex v_i in turn independently infects a Binomial$(N(v_i), p_1)$ number of its neighbors and so on. Once a neighbor is infected or is attempted to be infected, then it cannot be re-infected. In essence, we produce a random number of offspring at each step to emulate the power-law degree distribution of the graph, and then we *thin* the offspring to emulate the selective retweeting process.

We mentioned before that the motivation for this cascade model is the shape of cascades on Twitter. Let us digress a moment to explain this in more detail.

Ordinary branching processes cannot generate the type of the cascades we observe in real life. We mentioned before that a prominent feature of cascades on social networks such as Twitter is that many of them look like giant star graphs—i.e. at distance 1 from the

source, the number of individuals is huge, but there are 0 or very few individuals past that. However, it is well known that:

Theorem 5.2.1. *Let T be the hitting time to 0 of a branching random walk associated with a branching process with offspring X having mean $\mu = \mathbb{E}\, X > 1$. Then:*

$$\mathbb{P}(k \leqslant T < \infty) \leqslant \frac{e^{-Ik}}{1 - e^{-I}}$$

where the exponential rate $I = \sup_{t \geqslant 0}(t - \log \mathbb{E}(e^{tX})) > 0$.

This suggests that when the branching process is huge, then it survives with high probability. In other words, we should not be able to observe many star graphs in social network cascades if they behave like ordinary branching processes, because any branching process which starts out as a huge star graph tends to survive long past the 1st generation. Our thinning setup is a natural way to reconcile this.

As opposed to the study of preferential attachment, we are not so much interested in how the underlying graph came to have a power law distribution as we are in the cascade dynamic that happens upon such a graph. Therefore we simply suppose G has a power-law degree distribution in the sense that G follows a configuration model where the degree sequence are iid random variables following a power-law distribution.

Our first goal is to show that the exploration of the neighborhood of a fixed node can be well-approximated by a branching process, and if the power law exponent is $\in (2, 3)$, then the approximating branching process has infinite mean. To do this we follow [107] and use a variant of the classic configuration model with fixed degrees. Fix an integer n and let $D_1, \ldots, D_n$ be iid copies of a generic random variable D with pmf obeying a power law

$$\mathbb{P}(D = k) := f(k) \propto k^{-\alpha}, \quad \alpha \in (2, 3),\ k \in \mathbb{Z}^+$$

We construct a graph on n vertices where vertex i has degree D_i $(i = 1, \ldots, n)$ in the way of the classic configuration model. Start with n vertices where each vertex i has D_i half-stubs. For each half-stub of which there are $\sum_{i=1}^{n} D_i = L_n$, we pair the stubs randomly to form edges between vertices. More specifically, for the first stub, pick one of the $L_n - 1$ remaining stubs uniformly at random and pair the two. Then pick a remaining unpaired stub and continue the process until all stubs are connected. Henceforth we shall refer to do this model as the $G(n, f)$ model.

Despite the degree distribution random variable D having $\alpha \in (2, 3)$, the number of neighbors $N(v_i)$ in a local exploration around an arbitrary node v_0 has infinite mean due to size-biasing. Suppose we begin with a node v_0 and want to investigate the number of neighbors of v_1, a neighbor of v_0. That is, we would like to know how many stubs are attached to v_1 not counting the one attached to v_0. Conditional on knowing $D_1, \ldots, D_n$, the probability that the number of other stubs $= k$ is approximately equal to the probability that a randomly chosen stub from all L_n stubs (i.e. the one attached to v_0) is attached to a node with $k + 1$ stubs. This is easily seen to be

$$\hat{f}^{(n)}(k) = \frac{k+1}{L_n} \sum_{j=1}^{n} \mathbb{1}(D_j = k + 1)$$

If $\alpha \in (2, 3)$ then the mean $\mathbb{E}\, D$ exists so that by the law of large numbers,

$$\hat{f}^{(n)}(k) \to \frac{(k+1)f(k+1)}{\mathbb{E}\, D} := \hat{f}(k)$$

This is a size-biased version of the initial degree distribution which now has infinite mean. If the graph is "large enough" so that we may ignore dependence from cycles and finite graph size, then we can continue this argument assuming independent stub choices each time so that the growth of the explored cluster is given by a branching process with offspring distribution $\{\hat{f}(k)\}_{k \geqslant 0}$. Note carefully that the size-biasing argument begins from the *second*

generation of the branching process onwards. The degree of an vertex chosen uniformly at random follows the distribution f as designed, and the root is chosen in such a way.

All in all we see that a key tool we will need are branching processes with infinite mean. But more than that, to run the cascade process on the graph we will need to **thin** the infinite-mean size-biased distribution $\hat{f}$ according to a Binomial distribution with probability decreasing with the distance from the root. Crucially, this resulting offspring distribution of the *cascade* will still have infinite mean:

Lemma 5.2.2. *Let Y be a discrete power-law random variable with exponent $\alpha \in (1,2)$:*

$$\mathbb{P}(Y = k) = k^{-\alpha}\frac{1}{\zeta(\alpha)}, \quad \alpha \in (1,2), k \geqslant 1$$

And let $X = Binomial(Y, p)$ for $p \in (0,1)$. Then $\mathbb{E}\, X = \infty$.

Proof:

Set $p_k = \mathbb{P}(X = k)$. By the law of total probability, we need to check that:

$$\sum_{k\geqslant 0} kp_k = \sum_{k\geqslant 0}\left[k\sum_{n\geqslant k}\binom{n}{k}p^k(1-p)^{n-k}\frac{1}{n^\alpha\zeta(\alpha)}\right] = \infty$$

It will be sufficient to show the bound $kp_k \geqslant C/(k-1)$ for $k \geqslant 2$:

$$\begin{aligned}
kp_k &\propto \sum_{n\geqslant k} k\binom{n}{k}p^k(1-p)^{n-k}\frac{1}{n^\alpha} \\
&= p\sum_{n\geqslant k-1}\binom{n}{k-1}p^{k-1}(1-p)^{n-(k-1)}\frac{n}{n^\alpha} \\
&= \frac{p^2}{k-1}\sum_{n\geqslant k-2}\binom{n}{k-2}p^{k-2}(1-p)^{n-(k-2)}\frac{n(n-1)}{n^\alpha} \\
&\geqslant \frac{p^2}{k-1}\underbrace{\sum_{n\geqslant k-2}\binom{n}{k-2}p^{k-2}(1-p)^{n-(k-2)}}_{1/p}
\end{aligned}$$

■

This fact combined with the decreasing probabilities of the thinning mechanism brings us into the new territory of infinite-mean **varying-environment** branching processes.

5.2 5.2.3. Coupling to a graph: a sketch

The motivation of our study is a graph cascade, but the focus of our study will be a branching process. Before we jump off into branching process land, let us make a quick note about how one might couple the two processes to translate results between the two domains.

As alluded to before, given a $G(n, f)$ configuration model graph and an arbitrary starting node v_0, the decreasing cascade model with probability sequence $\mathbf{p} = \{p_n\}$ will trace out a tree starting from that node. As argued before, locally to v_0 the growth of the tree should be well-approximated by an infinite-mean, decreasing BP.

More specifically a first step is to show the following. Given a $G(n, f)$ graph and a root node v_0 picked uniformly at random let the decreasing cascade tree at time N starting from v_0 be denoted $T^f_{\mathbf{p}}(N)$.

As argued above, local to a randomly chosen node the exploration of $G(n, f)$ is roughly approximated by a branching process with the size-biased offspring distribution $\hat{f}$. In fact as argued above, we actually need a *delayed* branching process where the offspring distribution for the first generation is f, then switching to $\hat{f}$ for all future generations.

To emulate the cascade, let D_f and $D_{\hat{f}}$ be generic independent random variables with distributions f and $\hat{f}$, respectively. Then let the *sequence* of distributions $\mathbf{h} = \{h_n : n \geqslant 1\}$ be defined by

$$h_1 \sim \text{Binomial}(D_f, p_1), \quad \text{and} \quad h_k \sim \text{Binomial}(D_{\hat{f}}, p_k),\ k \geqslant 2$$

Let $Z_{\mathbf{h}} = \{Z_{\mathbf{h}}(N), N \geqslant 1\}$ denote a branching process with varying environment at time N with offspring distributions given by $\mathbf{h}$, viewed as a tree. Except for the first generation, $Z_{\mathbf{h}}$ is an infinite-mean branching process with varying environment

This branching process is easier to deal with than the general cascade process on the graph $G(n, f)$. To couple the branching process to the graph cascade, one simply needs to show that:

Proposition 5.2.3. *For fixed time N, as the size of the graph $G(n, f) \to \infty$ then:*

$$\lim_{n\to\infty} \mathbb{P}(Z_{\mathbf{h}}(N) = T^f_{\mathbf{p}}(N)) = 1$$

Once this is established, then the problem reduces essentially to a study of branching processes. For brevity, we will not carry out the rest here (see [107] for the proof which inspired this).

5.3 5.3. Analysis of the branching process

5.3 5.3.1. The thinned branching process

To study the behavior of these branching processes, we follow the approach of [94] and [58]. To distinguish between the general branching process setting and the binomial setting, we shall refer to the binomially-pruned branching process as an *thinned branching process* from now on, which will be totally determined in our case by the exponent of the power law distribution $\alpha \in (1, 2)$ and the *thinning probabilities* $\{p_k\}_{k\geqslant 1}$.

With non-varying offspring distributions, infinite-mean processes are just supercritical Galton-Watson processes which have positive probability of survival. But it is not clear what happens when the offspring distribution is progressively thinned. We might suspect that thinned branching processes separate into subcritical and supercritical regimes similar to classical branching processes, depending on how quickly the attenuation probabilities $\{p_k\}$

go to 0. Therefore our first task is to determine whether a guaranteed-extinction regime even exists for this thinned process.

In the rest of the section we will use the following notation:

Definition 5.3.1. *Let $\{Z_n\}_{n\geqslant 0}$ denote a thinned branching process driven by offspring distribution $\{g\}$ and thinning probabilities $\{p_n\}_{n\geqslant -0}$ and let X_i^n denote individual i in generation n for $1 \leqslant i \leqslant Z_{n-1}$. The process evolves as follows:*

1. *Each ith individual in the nth generation X_i^n has a random number of children W_i^n according to a distribution with p.g.f. g.*

2. *W_i^n is thinned so that only $Y_i^n \sim Binomial(W_i^n, p_n)$ children survive to the $(n+1)$th generation.*

We start the process at $Z_0 = 1$ so that the number of individuals at generation $n = 1$ is given by X_1^0.

We will make extensive use of generating functions, but we will need to distinguish between the pre- and post-thinning distributions.

Definition 5.3.2. *We define three sets of pgfs relating to an individual X_i^n in the nth generation of the branching process:*

1. *Let $f_n(s)$ be the pgf for the post-thinning, **final** number of offspring Y_i^n of X_i^n.*

2. *Let $g(s)$ be the pgf for the pre-thinning, **initial** number of children W_i^n of X_i^n.*

3. *Let $F_n(s)$ be the pgf for Z_n, the total number of individuals in generation n.*

Throughout, we will use the phrase "offspring distribution" to refer to the offspring distribution's pgf g and "thinned distributions" to refer to the distributions corresponding to the pgfs $\{f_n\}_{n\geqslant 0}$.

*Note that $F_0(s) = s$ and that f_0 is **not** identity as it is the generating function of the number of offspring of X_1^0.*

Assumption. *We assume throughout that $f_n'(1) = \infty$ for all n so that the mean number of offspring in each generation is infinite.*

Note that $F_n = f_0 \circ f_1 \circ \cdots \circ f_{n-1}$, and $f_n(s)$ is infinitely differentiable on $(0, 1)$ and left-continuous at $s = 1$.

As it turns out, calculating the extinction probability by evaluating the n-fold composition is difficult given the types of offspring distributions in our situation. Consider our previous example of offspring distributions with form Binomial(X, p) where X is an infinite-mean, $\mathbb{Z}^+$-valued power-law random variable with tail exponent α. That is, X obeys $\mathbb{P}(X > k) \propto k^{-\alpha}$ for $k \in \mathbb{Z}^+$ where $\alpha \in (0, 1)$. Then the generating function of the distribution after Binomial thinning $f(s, p, \alpha)$ is

$$\begin{aligned} f(s, p, \alpha) &= \mathbb{E}\left[\mathbb{E}(t^{\mathrm{Bin}(x,p)} | X = x)\right] \\ &= \zeta(\alpha + 1)^{-1} \cdot \sum_{k=1}^{\infty} k^{-\alpha+1}(1 - p + ps)^k \\ &= \frac{\mathrm{Li}_{\alpha+1}(1 - p + ps)}{\zeta(\alpha + 1)} \end{aligned} \tag{5.1}$$

Where $\mathrm{Li}_\alpha(s)$ is the polylogarithm of order $\alpha \in (0, 1)$ and $\zeta(\cdot)$ is the Riemann zeta function. These are not nice functions to deal with. Our proof of the extinction criteria will therefore lean on an approximation for $f(s, p, \alpha)$ for s close to 1.

5.3 5.3.2. The extinction criteria

As alluded to above, the intuition is that if the offspring distributions are thinned "fast enough," then it should be possible to force the process to extinguish. Let $q := \mathbb{P}(\exists n : Z_n = 0)$. So what is "fast enough?" The main result of this section is

Theorem 5.3.3. *Suppose that G is a probability distribution satisfying*

$$1 - G(x) \sim \frac{C}{x^\alpha}, \quad \alpha \in (0, 1) \tag{5.2}$$

Then a branching process with binomial thinning of such an offspring distribution extinguishes with probability 1 if and only if the thinning probabilities $\{p_n\}_{n\geqslant 1}$ *satisfy*

$$-\sum_{k=1}^{n}(1/\alpha)^{-k}\log p_k \to \infty \quad as\ n \to \infty$$

Clearly sequences which satisfy $p_n^{\alpha^n} \to 0$ satisfy this criterion, examples of which include:

1. $p_n = \exp(-n\alpha^{-n})$
2. $p_n = n^{-\alpha^{-n}}$

To see right off the bat why 5.3.3 might hold, note that in [41] it is shown that for certain well-behaved infinite-mean branching processes *without* thinning there exists a constant $\beta \in (0,1)$ such that $\beta^n(\log Z_n + 1)$ converges a.s. to a random variable which is finite with positive probability. Essentially this says that the unthinned process Z_n grows at rate $e^{\beta^{-n}}$, so the critical threshold for the rate of thinning should be roughly the same.

To be slightly more precise, we might try the classical way to study extinction probabilities by way of the recursion

$$F_n(0) = f_0(f_1(\cdots f_{n-1}(0)))$$

where the limit $\lim_{n\to\infty} F_n(0) = \mathbb{P}(\exists n : Z_n = 0)$. Therefore the question of guaranteed extinction boils down to the question of the limit of successive compositions of pgf's. Fortunately this has been investigated in the general case in [34] who showed the following key result:

Theorem 5.3.4. *The sequence* $\{F_n\}_{n\geqslant 0}$ *converges to a limit* F *defined by* $F(1) = 1$ *and* $F(s) = \sum_{k=0}^{\infty} b_k s^k$ *for* $0 \leqslant s < 1$, *where each* $b_k \geqslant 0$ *and* $\sum_{k=0}^{\infty} b_k \leqslant 1$.

Proof: Theorem 2, [34] or [8].

Note that the limit F satisfies $F(0) = b_0$ for some $b_0 \in [0,1]$ and also $F(1-) = \sum_{k=0}^{\infty} b_k$. In other words, while each f_k is left-continuous at 1, this may not be the case for the limit of the compositions. We have $0 \leqslant F(0) \leqslant F(1-) \leqslant 1$, so to settle whether $F(0) = 1$ we have to check when $F(0) = F(1-)$ and $F(1-) = 1$. Unfortunately easy conditions for checking these in general don't exist. Instead, we rely on clever vectorized approach from [50].

First we need some vector notation. Let I^{∞} be the space of infinite sequences with values in $[0,1]$. Let "$<$" denote the usual coordinate-wise operation on sequences, and for real sequences $\bar{a} := \{a_k\}_{k\geqslant 0}$ and $\bar{b} := \{b_k\}_{k\geqslant 0}$ with $b_k \neq 0$, let $\bar{a}/\bar{b} = \{a_k/b_k\}_{k\geqslant 0}$, $\bar{a} - \bar{b} = \{a_k - b_k\}_{k\geqslant 0}$ and $c \cdot \bar{a} = \{c \cdot a_k\}_{k\geqslant 0}$ for $c \in \mathbb{R}$. Furthermore let $\mathbf{1} = \{1, 1, \ldots\}$.

Denote by $Z(k, k+j)$ the number of individuals in the $(k+j)$th generation of the branching process conditional on there being exactly one individual in the system at generation k. Write

$$e_k = \mathbb{P}(Z(k, k+j) = 0 \text{ for some } j > 0$$

and let $\bar{e} = \{e_k\}_{k\geqslant 0}$. Note that either $\bar{e} = \mathbf{1}$ or $\bar{e} < \mathbf{1}$. Now define the function $\bar{f} : I^{\infty} \to I^{\infty}$ by

$$\bar{f}(\bar{s}) = \{f_k(s_{k+1})\}_{k\geqslant 0}, \quad \bar{s} = \{s_1, s_2, \ldots\} \in I^{\infty}$$

Then we have the main result of [50]:

Theorem 5.3.5. *For any inhomogeneous branching processes defined by the sequence of offspring distributions $\{f_k\}_{k\geqslant 0}$, $\bar{e} < 1$ if and only if there exists $\bar{s} \in I^{\infty}, \mathbf{0} \leqslant \bar{s} < \mathbf{1}$ such that,*

$$\bar{f}(\bar{s}) \leqslant \bar{s} \tag{5.3}$$

This theorem will be the basis for our proof.

5.3 5.3.3. Proof of Theorem 5.3.3

We can make the condition of Theorem 5.3.5 a little easier to check for our case by observing that our thinning process narrows down the types of sequences $\bar{s}$ we need to consider.

In the context of our thinned branching process, the requirement that $p_n \downarrow 0$ implies that the sequence $\bar{s}$ must satisfy $s_n \to 1$. It also should not be surprising that any finite number of leading terms of the sequence (s_n) don't matter so that it is enough for us to check this condition as $n \to \infty$. Let us summarize these arguments in a lemma:

Lemma 5.3.6. *For a thinned branching process driven by thinned distributions with pgfs* $\{f_n\}_{n\geqslant 0}$, $e < 1$ *if and only if there exists a sequence* $\{s_n\}_{n\geqslant 0}$, $s_n \in [0,1)$ *and* $s_n \to 1$ *such that for some* N

$$f_n(s_{n+1}) \leqslant s_n, \quad \forall n \geqslant N \tag{5.4}$$

Proof:

We begin by arguing that there is no loss of generality in considering only $\{s_n\}$ with $s_n \to 1$. If there does in fact exist $\bar{s}$ such that $\bar{f}(\bar{s}) \leqslant \bar{s}$, then it is necessary that $s_n \to 1$. To see why, note that $f_n(s) \geqslant f_n(0)$ for all $s \in [0,1]$. Therefore if $f_n(s_{n+1}) \leqslant s_n$ for all n, then $s_n \geqslant f_n(0)$ for all n. Since $p_n \to 0$, then $f_n(0) \to 1$ so $s_n \to 1$ also.

The fact that $s \to 1$ is necessary also implies that if there does not exist $\bar{s}$ satisfying $\bar{f}(\bar{s}) \leqslant \bar{s}$ with $s_n \to 1$, then there cannot exist any other such sequence $\tilde{s} \in I^\infty$ satisfying $\bar{f}(\tilde{s}) \leqslant \tilde{s}$.

To see how the tail condition 5.4 implies the full condition on $\bar{s}$ in Theorem 5.3.5, note that we can start from s_N and then construct the initial elements of the sequence $s_1, \ldots, s_{N-1}$ satisfying condition 5.3 by setting $s_{N-1} = f_{N-1}(s_N) \in (0,1)$. ■

Now letting $g(s,\alpha)$ be the pgf of the offspring distribution, the thinned distribution $f(s,p,\alpha)$ for a fixed thinning probability p may be written $g(1-p+ps,\alpha)$ so that the survival criteria in Lemma 5.3.6 is that there exists $\{s_n\}_{n\geqslant 1}$, $s_n \in [0,1)$ for all n with $s_n \to 1$

such that

$$g(1 - p_n + p_n s_{n+1}, \alpha) \leqslant s_n, \quad \text{for all } n$$

It will be convenient to work on a simpler approximating function for $g(\cdot)$. For future reference we introduce the notation

$$h(s) := (-\log(s))^\alpha \tag{5.5}$$

The approximation we will work with is

$$g(1 - p + ps) \sim 1 - C \cdot h(1 - p + ps) = 1 - C(-\log(1 - p + ps))^\alpha$$

holding for all $C > 0$ as $s \to 1$. Clearly, any function converging to 0 can be used in place of h for this approximation, but the reason for our particular choice will become clear shortly.

The survival criteria is now

$$1 - Ch(1 - p_n(1 - s_{n+1})) \leqslant s_n, \quad \text{some sequence } s_n \to 1$$

To establish the possibility of $(s_n)_{n \geqslant 0}$ satisfying this criteria, we will make use of the following deterministic lemma:

Lemma 5.3.7. *For brevity, say that $(p_n)_{n \geqslant 0}$ satisfies $(\star)$ if there exists $b \in (0, 1)$ such that:*

$$-\sum_{k=1}^{n} \gamma^{-k} \log p_k \to -\log b$$

Fix some $\gamma > 1$ and suppose that $(p_n)_{n \geqslant 0}$ is a sequence $\in (0, 1)$ which converges to 0. Note: in what follows below the sequences (a_n) depend on the constant C but we suppress this notation for simplicity.

1. *If $(p_n)_{n \geqslant 1}$ satisfies $(\star)$, then for **every** $C > 0$ there exists a sequence $(a_n)_{n \geqslant 0} \in (0, 1]$ also converging to 0 such that $p_n \geqslant C \frac{a_n^\gamma}{a_{n+1}}$ for all $n \geqslant 1$.*

2. *If for* ***some*** $C > 0$ *there exists a sequence* $(a_n)_{n\geqslant 0} \in (0,1]$ *also converging to 0 such that* $p_n \geqslant C\frac{a_n^\gamma}{a_{n+1}}$ *for all* $n \geqslant 1$*, then* $(p_n)_{n\geqslant 1}$ *satisfies* $(\star)$.

Proof: For the direct portion we show that a sequence $(a_n)_{n\geqslant 1}$ satisfying the theorem can be constructed. Observe that for the $(n+1)$th term a_{n+1} to satisfy the stated condition, we must have:

$$a_{n+1} \geqslant C\frac{a_n^\gamma}{p_n}$$

Iterating this n times we see that

$$a_{n+1} \geqslant C^{\gamma+\gamma^2+\ldots+\gamma^n}\frac{a_1^{\gamma^n}}{\prod_{k=1}^n p_k^{\gamma^{(n-k)}}}$$

Now assume without loss of generality that $C < 1$ so that the term involving C is bounded above by 1 (if $C \geqslant 1$ then it can be absorbed into the p_k's by $p_k \mapsto p_k/C$ and the rest of the argument goes through). If we show that $a_1^{\gamma^n}/\prod_{k=1}^n p_k^{\gamma^{(n-k)}} \to 0$, then this guarantees that we will be able to pick (a_n) satisfying the stated inequality with $a_n \in (0,1)$ for all n and $a_n \to 0$.

Take logs and set the initial value $a_1 = b - \epsilon$ where $\epsilon \in (0, b)$. Then

$$\begin{aligned}\log\frac{a_1^{\gamma^n}}{\prod_{k=1}^n p_k^{\gamma^{(n-k)}}} &= \gamma^n \log b - \sum_{k=1}^n \gamma^{(n-k)}\log p_k \\ &= \gamma^n\left(\log b - \sum_{k=1}^n \gamma^{-k}\log p_k\right)\end{aligned}$$

Since $\sum_{k=1}^n \gamma^{-k}\log p_k \to \log b$ then the quantity inside the parenthesis eventually stays negative and the entire RHS $\to -\infty$, which implies that the original ratio $\to 0$.

For the converse portion, suppose that the contrapositive is not true. That is, suppose $-\sum_{k=1}^n \gamma^{-k}\log p_k \to \infty$ but that there exists $(a_n)_{n\geqslant 0}$ such that

$$p_k \geqslant C \cdot \frac{a_k^\gamma}{a_{k+1}} \quad \text{for all } k \geqslant 1$$

Rearranging this we have the relation

$$a_k \leqslant \left(\frac{p_k a_{k+1}}{C}\right)^{1/\gamma}$$

Now starting at a_1 and expanding the right-hand side recursively n times we obtain

$$a_1 \leqslant \frac{(p_1^{\gamma^{-1}} p_2^{\gamma^{-2}} \cdots p_n^{\gamma^{-n}}) a_{n+1}^{\gamma^{-n}}}{C^{\gamma^{-1}+\gamma^{-2}+\cdots+\gamma^{-n}}}$$

Since $C < 1$ then $\gamma^{-1} + \cdots + \gamma^{-n} \leqslant \gamma/(\gamma - 1)$ for all n so the denominator is bounded above by $C_0 = C^{\gamma/(\gamma-1)}$. If $C > 1$ then this quantity is bounded by $C_0 = C^{1/\gamma}$. Combining this with the fact that $a_n < 1$ for all n, the inequality simplifies to

$$a_1 \leqslant \frac{\prod_{k=1}^n p_k^{\gamma^{-k}}}{C_0}$$

However, $-\sum_{k=1}^n \gamma^{-k} \log p_k \to \infty$ implies $\prod_{k=1}^n p_k^{\gamma^{-k}} \to 0$, thus sending $n \to \infty$ above implies $a_1 = 0$ and we have a contradiction. ■

From there it is only a short hop to the result for our approximation sequence:

Lemma 5.3.8. *Again for brevity, say that $(p_n)_{n\geqslant 0}$ satisfies $(\star)$ if there exists $b \in (0,1)$ such that:*

$$-\sum_{k=1}^{n} \gamma^{-k} \log p_k \to -\log b$$

1. *If $(p_n)_{n\geqslant 1}$ satisfies $(\star)$, then for **every** $C > 0$ there exists a sequence $(s_n)_{n\geqslant 0} \in [0,1)$ converging to 1 such that $1 - C \cdot h(1 - p_n(1 - s_{n+1})) \leqslant s_n$ for all $n \geqslant 1$.*

2. *If for **some** $C > 0$ there exists a sequence $(s_n)_{n\geqslant 0} \in [0,1)$ converging to 1 such that $1 - C \cdot h(1 - p_n(1 - s_{n+1})) \leqslant s_n$ for all $n \geqslant 1$, then $(p_n)_{n\geqslant 1}$ satisfies $(\star)$.*

Proof:

It is sufficient to show the relevant properties for sequences $(s_n)_{n\geqslant 0}$ satisfying the easier inequality

$$1 - C \cdot \Big[p_n(1 - s_{n+1})\Big]^{\alpha} \leqslant s_n \text{ for all } n \geqslant 1 \tag{5.6}$$

since, by the identity $1 - x \leqslant -\log(x),\ x > 0$ and the facts that $p_n, s_n \in [0, 1)$ the following inequality holds for all n:

$$1 - C \cdot h(1 - p_n(1 - s_{n+1})) \leqslant 1 - C \cdot \Big[p_n(1 - s_{n+1})\Big]^{\alpha}$$

Inequality (5.6) implies that we want to study sequences $(s_n)_{n\geqslant 0}, s_n \in [0, 1)$ for all n such that:

$$p_n \geqslant C \cdot \frac{(1 - s_n)^{1/\alpha}}{1 - s_{n+1}}, \quad \text{for all } n \geqslant 1 \tag{5.7}$$

Now applying the previous lemma with the map $1 - s_n = a_n$ gives the result. ■

Finally we need to tighten the result above concerning the approximating function $h(\cdot)$ to the pgf $g(\cdot)$. The key tool is a Tauberian theorem giving the behavior of the Laplace-Stieltjes transform of a heavy-tailed random variable.

Theorem 5.3.9. *([16], Theorem 8.1.6) Let $F(\cdot)$ be the CDF of some probability distribution and $\phi(\cdot)$ its Laplace-Stieltjes transform. Then for $0 \leqslant \alpha < 1$ and a slowly-varying function ℓ, the following are equivalent*

$$(1) \quad 1 - \phi(s) \sim s^{\alpha}\ell(1/s), \quad s \downarrow 0$$
$$(2) \quad 1 - F(x) \sim \frac{\ell(x)}{x^{\alpha}\Gamma(1 - \alpha)}, \quad x \to \infty$$

The relationship between the Laplace-Stieltjes transform $\phi(\cdot)$ and the probability generating function $g(\cdot)$ is $\phi(-\log s) = f(s)$. Also, for our offspring distributions the function $\ell(x)$ is a constant $C_{\alpha} > 0$ for all $\alpha \in (0, 1)$. So making the map $s \mapsto -(\log s)$, we can rewrite $\phi(-\log s)$ as our pgf $g(s)$ and we have the result (now for $s \to 1$):

Corollary 5.3.10. *Let $g(\cdot)$ be the pgf of an offspring distribution satisfying 5.2 from Theorem 5.3.3. Then there exists $C > 0$ such that*

$$1 - g(s) \sim C \cdot h(s), \quad s \to 1$$

Where $h(\cdot)$ is defined in 5.5.

This implies the following simple Lemma:

Lemma 5.3.11. *Given an offspring distribution satisfying 5.2, then there exist constants $C_0, C_1 > 0$ such that*

1. *There exists s_0 such that $1 - g(s) \leqslant C_0 \cdot h(s)$ for all $s > s_0$*
2. *There exists s_1 such that $1 - g(s) \geqslant C_1 \cdot h(s)$ for all $s > s_1$*

Proof: Solving (1) for C_0 gives

$$\frac{1 - g(s)}{h(s)} \leqslant C_0 \quad \text{for all } s > s_0$$

Now by Corollary 5.3.10, there exists $C > 0$ such that $(1 - g(s))/h(s) \to C$. Thus setting $C_0 > C$ implies the result. The proof for (2) is exactly analogous. ■

Combining Lemma 5.3.11 and Lemma 5.3.8, extracting the $\{p_n\}_{n \geqslant 0}$ criterion from the proof and making the change $\gamma = 1/\alpha > 1$ completes the proof of Theorem 5.3.3.

CHAPTER 6

Future directions

We discuss some potential extensions of the two works presented above, and some ideas for new work in unrelated areas.

6.1 6.1. Changepoint

The methodology developed in Chapter 1 is sufficiently general to apply to several extensions of the preferential attachment model studied there.

6.1 6.1.1. Timing

A straightforward question arising from the current work is the question of the timing of the changepoint. Recall that in our current setup the changepoint occurs at a time γn, $\gamma \in (0, 1)$, in other words a fixed fraction of the time. In this case we have shown that the changepoint does not occur early enough in the process to pick up any effect from the post-change attachment regime (switching from $+\alpha$ to $+\beta$ in the attachment probabilities). It stands to reason then that if we push the changepoint earlier in the process to a time n^γ, $\gamma \in (0, 1)$ then the process *will* feel the effect of the change.

6.1 6.1.2. Non-linear attachment

Yet another potentially more interesting direction for investigation involves variations of the generative process itself. Recall that the essence of the model is attachment of new vertices to existing ones with probability proportional to the existing degree of the vertex.

Mathematically, if v is an existing vertex in the graph at time n and $D(v,n)$ is the degree of v at time n, then the probability that a new vertex connects to v at time n is some function f of $D(v,n)$:

$$\mathbb{P}(\text{connect to } v) \propto f(D(v,n)) \tag{6.1}$$

In the model of chapter 1 (*linear preferential attachment*) this function has the form $f(v,n) = D(v,n) + \alpha$, $\alpha > -1$. There are a couple possible ways to generalize this model.

The most straightforward extension is to generalize the function f to, for example, an arbitrary positive increasing function or some distribution function on the positive integers. This has echoes of the classical problem of [27] but retains the strong dependence structure of preferential attachment.

Another natural extensions involves introducing some extra randomness to the model in a way called *preferential attachment with fitness*. In this case we suppose there is a fixed probability distribution λ on $\mathbb{R}^+$ and each node added to the graph is born with random fitness g_v independently drawn from λ and new vertices connect to v with probability depending on this fitness:

1. Preferential attachment with additive fitness: $f(D(v,n)) = D(v,n) + g_v$
2. Preferential attachment with multiplicative fitness: $f(D(v,n)) = g_v \cdot D(v,n)$

The case of additive fitness is essentially linear preferential attachment with random additive constant. Extending the changepoint estimation problem to these setups is a natural next step: suppose that up to time γ the graph has fitness drawn from distribution λ_0 and after time γ the graph has fitness drawn from some other distribution λ_1. Detecting the changepoint in this case combines the theory developed in our project with the classical theory of univariate changepoint detection in an independent sample.

6.1 6.1.3. Preferential attachment with types

Suppose we have a preferential attachment-type dynamic network model with the following wrinkle. Suppose that new nodes entering the network are of one of two types. If a node is of type 0, it attaches to existing nodes with probability proportional to some function f of the node degree. If a node is of type 1, it attaches to existing nodes with probability proportional to a possibly different function g of the node degree.

More precisely, let G_n be a graph on n vertices grown according to preferential attachment and let $G_n(t)$ denote the view of the graph at time t, corresponding to the instant when the graph had tn vertices.

Definition 6.1.1. *On a preferential attachment graph G_n, let $a(i) \in \{0, 1\}$ be the type of the ith vertex in the graph.*

The basic question: Given a graph grown by this scheme of size n, when can we successfully test

$$H_0 : f = g \quad \text{vs.} \quad H_A : f \neq g$$

This has, as a special case, our ordinary changepoint problem. Just define the types as

$$a(i) = \begin{cases} 0, & i \leqslant \gamma n \\ 1, & i > \gamma n \end{cases} \tag{6.2}$$

In light of this, there are two first steps to investigate.

1. Stay in the changepoint regime of (1) and study the effect of general f and g. One can think of this as an even further, nonparametric extension of the non-linear attachment case set out in Section 6.1.2.

2. Keep $f = d(v) + \alpha$ and $g = d(v) + \beta$, but study the effect of $a(i)$. For example, how does the following scheme differ from ordinary linear preferential attachment at all?

$$a(i) = \begin{cases} 0, & i \text{ odd} \\ 1, & i \text{ even} \end{cases}$$

Or, letting $\{X_i\}_{i=1,\ldots,n}$ be an iid sequence of Bernoulli(p) random variables,

$$a(i) = X_i$$

In the simplest cases of linear or non-linear preferential attachment, the notion of *types* is deliberately meant to invoke the theory of continuous general branching processes with types set out in [65]. Indeed, the theory for the asymptotic growth behavior of such processes mirror the theory in Chapter 3 quite closely. As a first step, such an extension should not be difficult.

From here, there are many different follow-up questions we can ask. Can we extend this to multiple "types" e.g. $f_1, f_2, \ldots f_k$. What if we do not know k? How would we get an estimate of this k? Further how would we cluster nodes into different types based on data? Now run our algorithm on citation networks (e.g. physics). Do the types that we get correlate with known communities or known clusters found by community detection algorithms? If not what new features do these produce?

6.2 6.2. Cascades

There are three natural directions in which the decreasing cascade research can be taken in.

6.2 6.2.1. The growth of the supercritical thinned branching process

The first is go down the road of Kesten-Stigum and try to study the growth of the thinned branching process. Let us flesh this out somewhat in detail.

It is possible to analyze the supercritical case of the infinite-mean branching process by using a martingale technique from [58]. Recall that if the mean (μ) is finite, then the growth rate for the branching process is essentially given by μ^n in the sense that Z_n/μ^n converges to an a.s. finite limit. In the infinite-mean supercritical case, it is known that no such normalizing sequence exists for Z_n unless Z_n is suitably normalized by some slowly-varying function $\ell(Z_n)$, e.g. $\log(Z_n + 1)$. The main thrust of this section is to investigate whether thinned branching processes behave more like the former or the latter.

To set this up we need to define some relatives of the generating functions.

Definition 6.2.1. *1. The cumulant generating function for the offspring of an individual from generation n (e.g. X_i^n) is:*

$$k_n(s) := -\log f_n(e^{-s}) = -\log \mathbb{E}(e^{-sX_n})$$

2. Also define the functional iterates of the cumulant generating functions as:

$$k^{(n)} = k_0 \circ k_1 \circ \cdots \circ k_{n-1}, \text{ where } k^{(0)}(s) = s$$

As with the pgf's, the iterated cumulant generating functions also turn out to be the cumulant generating functions for the number of individuals in generation n:

$$k^{(n)}(s) = -\log F_n(e^{-s}) = -\log \mathbb{E}[\exp(-sZ_n)]$$

The following properties about k_n will be important:

Lemma 6.2.2. *k_n is continuous, strictly increasing, and concave for $s \geqslant 0$. Furthermore $k_n(0) = 0$, $k_n'(0+) = \mathbb{E}\, X_n$ and $k_n(s) > s$ for $s \in (0, -\log q)$.*

It follows that these properties also hold for the iterates $k^{(n)}$. It follows that the inverses of k_n and $k^{(n)}$ are defined and have analogous properties:

Definition 6.2.3. *(The inverses of the cumulants)*

1. *Let the functional inverse of k_n be denoted*

$$h_n(s) = k_n^{-1}(s)$$

2. *Let the functional inverse of $k^{(n)}$, the iterate of k_n, be*

$$h^{(n)} = (k^{(n)})^{-1}, \text{ where } h^{(0)}(s) = s$$

And as expected, $h^{(n)}$ also satisfies $h^{(n)} = k_{n-1}^{-1} \circ k_{n-2}^{-1} \circ \cdots \circ k_0^{-1}$

These inverses are well-defined for all n in the interval $[0, -\log q)$, which exists only if the branching process has some probability of surviving indefinitely since $\lim_{s\to\infty} k^{(n)}(s) = -\log F_n(0)$ and $F_n(0) \uparrow q$. But as argued in the previous section, this is always the case. The martingale we need is from [58]:

Lemma 6.2.4. *Let $\{\mathcal{F}_n\}_{n\geqslant 0}$ be the filtration generated by $\{Z_n\}_{n\geqslant 0}$. By Theorem 5.3.3, $\mathbb{P}(\exists n : Z_n = 0) = q < 1$ so let $s \in (0, -\log q)$ and define the process*

$$M_n(s) := \exp\Big(- h^{(n)}(s) Z_n \Big) \tag{6.3}$$

Then $\{M_n(s)\}_{n\geqslant 0}$ is an $\mathcal{F}_n$-martingale.

Proof:

$$\begin{aligned}\mathbb{E}[M_n(s)|\mathcal{F}_{n-1}] &= \mathbb{E}\left[\exp(-h^{(n)}(s)Z_n)|\mathcal{F}_{n-1}\right] \\ &= \mathbb{E}\left[\exp(-h^{(n)}(s)X_1^{n-1})\right]^{Z_{n-1}} \\ &= \exp\Big(-k_{n-1}(h^{(n)}(s))Z_{n-1}\Big) \\ &= \exp\Big(-h^{(n-1)}(s)Z_{n-1}\Big)\end{aligned}$$

■Again, this martingale depends on the existence of inverses of the cumulant generating functions in a right neighborhood of zero, which is guaranteed in the supercritical case.

Straight away by the martingale convergence theorem and by boundedness of M_n, $M_n(s) \xrightarrow{\text{a.s.}} M(s)$ or equivalently,

$$Z_n/(h^{(n)}(s))^{-1} \xrightarrow{\text{a.s.}} W(s) := -\log M(s)$$

The formulation $W(s)$ of the martingale limit is of interest because, if $W(s)$ has positive probability in $(0,1)$, then the sequence $\{h^{(n)}(s))^{-1}\}_{n\geqslant 0}$ describes the growth of Z_n. Let us investigate this.

As a starting point, $\mathbb{E}\,M(s) = M_0(s) = e^{-s} < 1$. Next, we need to check whether or not $\mathbb{P}(0 < W(s) < \infty) = 0$. If it does, then $\mathbb{P}(W(s) \in \{0,\infty\}) = 1$ and the limit is not proper and non-degenerate. So we aim to show the proposition

Proposition 6.2.5. *With the notation above,* $\mathbb{P}(M(s) = 0) = \mathbb{P}(W(s) = \infty) < 1$.

Now if $\mathbb{P}(W(s) \in \{0,\infty\}) = 1$, then a normalizing sequence does not exist. Therefore following [94] we analyze whether $\mathbb{P}(0 < W(s) < \infty) > 0$.

As it turns out, whether or not this is the case will depend on the specific points s_0 at which the cgfs in the normalizing sequence $(h^{(n)}(s_0))^{-1}$ are evaluated. Borrowing the terminology of [94], suppose we say:

Definition 6.2.6. *A point $s \in (0, -\log q)$ is regular if $\mathbb{P}(W(s) \in \{0, \infty\}) = 1$ and irregular otherwise.*

Then a natural next step in deriving a Kesten-Stigum type result for the growth of these processes is to determine whether or not irregular points exist for thinned processes. This work is ongoing.

6.2 6.2.2. Shape

A second area of future research on cascades is to focus again on the shape of these cascades. Recall that our initial motivation for developing the theory of thinned branching processes was to understand the *shape* of real-life cascades. Ultimately, we would like to know whether the thinned process can be flexibly tuned to generate trees shaped like real cascades on Twitter or similar social media platforms. This ultimately will involve both empirical and theoretical work.

First, we need an understanding of the shape of real-world networks (this has already been done for *structural virality* in [53]). Secondly, we need to develop a strategy for studying the shape of branching process; not only do we need to understand the shape of the branching process probabilistically (e.g. the nth generation size conditioned on survival, the structural virality of the cascade by the nth generation), we also need to relate those behaviors to our specific choice of offspring distribution and the thinning sequence.

6.2 6.2.3. Inference for the virality of a cascade

The third direction is inferential. Assuming that our model is a reasonable approximation for the behavior of retweet cascades in reality, can we harness it to make predictions about, say, the final size of the cascade or the virality of the resulting cascade? There have been only a couple works to this regard, and all under the guide of predicting virality [111, 31].

The general idea as applied to our case is as follows: we might suppose that every tweet posted on Twitter has a "intrinsic" virality which is manifested in a thinning sequence on

the underlying follower graph. This is unobserved, but it is trivial to estimate it empirically: we simply calculate the empirical proportion of total outdegree at distance n from the source which retweeted the source. The general idea is to distinguish between content which is viral because it happens to go through a string of users with huge amounts of followers (thinning probabilities decay rapidly), and content which is viral because it is exceptionally interesting (thinning probabilities decay slowly).

Combining this empirical thinning sequence with our probabilistic knowledge of how a true thinning sequence impacts the final characteristics of a retweet cascade, we ought to be able to develop a simple inferential framework for predicting the virality of a tweet based on observing just the first few levels of the thinning sequence. To our knowledge, only [32] has attempted this in published research, and the author do not attempt to model the cascade itself. A framework incorporating an accurate model of such cascades would be promising. This work is closely tied up with the work of estimating influence probabilities in the independent cascade model, see Subsection 2.3.1 for references.

BIBLIOGRAPHY

[1] Aggarwal, C. C. and Yu, P. S. (2005). Online analysis of community evolution in data streams. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 56–67. SIAM.

[2] Akoglu, L., McGlohon, M., and Faloutsos, C. (2010). Oddball: Spotting anomalies in weighted graphs. In *Advances in knowledge discovery and data mining*, pages 410–421. Springer.

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

[4] Albert, R., Jeong, H., and Barabasi, A.-L. (1999). Diameter of the world-wide web. *Nature*, 401:130–131.

[5] Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J., and Tiwari, M. (2015). Global diffusion via cascading invitations: Structure, growth, and homophily. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 66–76. International World Wide Web Conferences Steering Committee.

[6] Athreya, K. B. and Karlin, S. (196812). Embedding of urn schemes into continuous time markov branching processes and related limit theorems. *Ann. Math. Statist.*, 39(6):1801–1817.

[7] Athreya, K. B. and Karlin, S. (1971a). Branching processes with random environments, ii: Limit theorems. *The Annals of Mathematical Statistics*, 42(6):1843–1858.

[8] Athreya, K. B. and Karlin, S. (1971b). On branching processes with random environments: I extinction probabilities. *The Annals of Mathematical Statistics*, 42(5):1499–1520.

[9] Bala, V. and Goyal, S. (1998). Learning from neighbours. *The review of economic studies*, 65(3):595–621.

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

[11] Basseville, M. and Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs.

[12] Bhamidi, S., Evans, S. N., and Sen, A. (2012). Spectra of large random trees. *Journal of Theoretical Probability*, 25(3):613–654.

[13] Bhamidi, S., Steele, J. M., and Zaman, T. (2014). Twitter event networks and the superstar model. *Accepted in Annals of Applied Probability*.

[14] Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026.

[15] Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M., and Wallach, D. S. (2015). Aggregate characterization of user behavior in twitter and analysis of the retweet graph. *ACM Trans. Internet Technol.*, 15(1):4:1–4:24.

[16] Bingham, N., Goldie, C., and Teugels, J. (1987). *Regular variation*. Cambridge University Press.

[17] Boas, R. (1977). Partial sums of infinite series, and how they grow. *American Mathematical Monthly*, pages 237–258.

[18] Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C., Gómez-Gardeñes, J., Romance, M., Sendina-Nadal, I., Wang, Z., and Zanin, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122.

[19] Boguña, M., Krioukov, D., and claffy, k. (2009). Navigability of complex networks. *Nature Physics*, 5:74–80.

[20] Bollobás, B. (2001). *Random graphs*. Cambridge University Press.

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

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

[23] Bollobás, B., Riordan, O., Spencer, J., and Tusnády, G. (2001-05). The degree sequence of a scale-free random graph process. *Random Structures and Algorithms*, 18(3):279–290.

[24] Brodsky, E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change point problems*. Springer Science & Business Media.

[25] Bubeck, S., Devroye, L., and Lugosi, G. (2016). Finding adam in random growing trees. *Random Structures & Algorithms*.

[26] Bubeck, S., Mossel, E., and Rácz, M. Z. (2015). On the influence of the seed graph in the preferential attachment model. *Network Science and Engineering, IEEE Transactions on*, 2(1):30–39.

[27] Carlstein, E. (1988). Nonparametric change-point estimation. *The Annals of Statistics*, pages 188–197.

[28] Carlstein, E. G., Müller, H.-G., and Siegmund, D. (1994). *Change-point problems*. IMS.

[29] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.

[30] Chen, W., Yuan, Y., and Zhang, L. (2010). Scalable influence maximization in social networks under the linear threshold model. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 88–97. IEEE.

[31] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936.

[32] Cheung, M., She, J., Junus, A., and Cao, L. (2016). Prediction of virality timing using cascades in social media. *ACM Trans. Multimedia Comput. Commun. Appl.*, 13(1):2:1–2:23.

[33] Chung, F. R. and Lu, L. (2006). *Complex graphs and networks*, volume 107. American mathematical society Providence.

[34] Church, J. (1971). On infinite composition products of probability generating functions. *Z. Wahrscheinlichkeitstheorie verw. Geb. 19*, pages 243–256.

[35] Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

[36] Cooper, C. and Frieze, A. (2003). A general model of web graphs. *Random Structures & Algorithms*, 22(3):311–335.

[37] Cowan, R. and Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of economic Dynamics and Control*, 28(8):1557–1575.

[38] Cozzo, E., Banos, R. A., Meloni, S., and Moreno, Y. (2013). Contact-based social contagion in multiplex networks. *Physical Review E*, 88(5):050801.

[39] Csörgö, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*, volume 18. John Wiley & Sons Inc.

[40] Curien, N., Duquesne, T., Kortchemski, I., and Manolescu, I. (2014). Scaling limits and influence of the seed graph in preferential attachment trees. *arXiv preprint arXiv:1406.1758*.

[41] Davies, P. L. (1978). The simple branching process: A note on convergence when the mean is infinite. *Journal of Applied Probability*, 15(3):466–480.

[42] Dodds, P. S. and Watts, D. J. (2004). Universal behavior in a generalized model of contagion. *Phys. Rev. Lett.*, 92:218701.

[43] Dorogovtsev, S. and Mendes, J. (2002). Evolution of networks. *Advances in physics*, 51(4):1079–1187.

[44] Dow, P. A., Adamic, L. A., and Friggeri, A. (2013). The anatomy of large facebook cascades. In *AAAI '13*.

[45] D'Souza, J. and Biggins, J. (1993). The supercritical galton-watson process in varying environments–the seneta-heyde norming. *Stochastic Processes and their Applications*, 48(2):237–249.

[46] Durrett, R. (2007). *Random graph dynamics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

[47] Durrett, R. and Resnick, S. I. (1978). Functional limit theorems for dependent variables. *The Annals of Probability*, pages 829–846.

[48] Eberle, W. and Holder, L. (2007). Discovering structural anomalies in graph-based data. In *Data mining workshops, 2007. icdm workshops 2007. seventh ieee international conference on*, pages 393–398. IEEE.

[49] Ethier, S. N. and Kurtz, T. G. (1986). *Markov processes: characterization and convergence*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

[50] Fearn, D. H. (1981). A fixed-point property for galton-watson processes. *Journal of Applied Probability*, 18(2):514–519.

[51] Fearn, D. H. et al. (1972). Galton-watson processes with generation dependence. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Health*. The Regents of the University of California.

[52] Goel, S., Anderson, A., Hofman, J., and Watts, D. J. (2016). The structural virality of online diffusion. *Management Science*, 62(1):180–196.

[53] Goel, S., Watts, D. J., and Goldstein, D. G. (2012). The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 623–638.

[54] Goldenberg, J., Libai, B., and Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 3(12):211–223.

[55] Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2010). Learning influence probabilities in social networks. *WSDM '10*.

[56] Harris, T. E. (2002). *The theory of branching processes*. Courier Corporation.

[57] Heard, N. A., Weston, D. J., Platanioti, K., Hand, D. J., et al. (2010). Bayesian anomaly detection methods for social networks. *The Annals of Applied Statistics*, 4(2):645–662.

[58] Heyde, C. (1970). Extension of a result of seneta for the super-critical galton-watson process. *The Annals of Mathematical Statistics*, 41(2):739–742.

[59] Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics reports*, 519(3):97–125.

[60] House, T. (2011). Modelling behavioural contagion. *Journal of The Royal Society Interface*, 8(59):909–912.

[61] Huang, Z. and Zeng, D. D. (2006). A link prediction approach to anomalous email detection. In *Smc*, pages 1131–1136.

[62] Iribarren, J. L. and Moro, E. (2009). Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.*, 103:038702.

[63] Iribarren, J. L. and Moro, E. (2011). Branching dynamics of viral information spreading. *Phys. Rev. E*, 84:046116.

[64] Jagers, P. (1974). Galton-watson processes in varying environments. *Journal of Applied Probability*, 11(1):174–178.

[65] Jagers, P. (1975). *Branching processes with biological applications*. Wiley.

[66] Jagers, P. and Nerman, O. (1984a). The growth and composition of branching populations. *Adv. in Appl. Probab.*, 16(2):221–259.

[67] Jagers, P. and Nerman, O. (1984b). Limit theorems for sums determined by branching and other exponentially growing processes. *Stochastic Process. Appl.*, 17(1):47–71.

[68] Jo, H.-H., Perotti, J. I., Kaski, K., and Kertész, J. (2014). Analytically solvable model of spreading dynamics with non-poissonian processes. *Physical Review X*, 4(1):011041.

[69] Juhasz, R., Kovacs, I. A., and Igloi, F. (2015). Long-range epidemic spreading in a random environment. *Physical Review E*, 91(3):032815.

[70] Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM.

[71] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on the World Wide Web*, pages 591–600. ACM.

[72] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., and Hurst, M. (2007). Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 551–556. SIAM.

[73] Liben-Nowell, D. and Kleinberg, J. (2008). Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638.

[74] Liptser, R. and Shiryayev, A. N. (2012). *Theory of martingales*, volume 49. Springer Science & Business Media.

[75] Liu, C. and Zhang, Z.-K. (2014). Information spreading on dynamic social networks. *Communications in Nonlinear Science and Numerical Simulation*, 19(4):896–904.

[76] Lyons, R., Pemantle, R., and Peres, Y. (1995). Conceptual proofs of l log l criteria for mean behavior of branching processes. *Ann. Probab.*, 23(3):1125–1138.

[77] MacPhee, I. and Schuh, H.-J. (1983). A galton-watson branching processes in varying environments with essentially constant offspring means and two rates of growth. *Australian & New Zealand Journal of Statistics*, 25(2):329–338.

[78] Marangoni-Simonsen, D. and Xie, Y. (2015). Sequential changepoint approach for online community detection. *Signal Processing Letters, IEEE*, 22(8):1035–1039.

[79] McCulloh, I. and Carley, K. M. (2011). Detecting change in longitudinal social networks. Technical report, DTIC Document.

[80] Móri, T. (2007). Degree distribution nearby the origin of a preferential attachment graph. *Electron. Comm. Probab*, 12:276–282.

[81] Newman, M. (2010). *Networks: an introduction*. Oxford University Press.

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

[83] Noble, C. C. and Cook, D. J. (2003). Graph-based anomaly detection. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining*, pages 631–636. ACM.

[84] Norris, J. R. (1998). *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. Reprint of 1997 original.

[85] Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925.

[86] Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, 84(14):3200.

[87] Peel, L. and Clauset, A. (2014). Detecting change points in the large-scale structure of evolving networks. *CoRR*, abs/1403.0989.

[88] Price, D. J. d. S. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

[89] Price, D. J. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5–6):292–306.

[90] Priebe, C. E., Conroy, J. M., Marchette, D. J., and Park, Y. (2005). Scan statistics on enron graphs. *Computational & Mathematical Organization Theory*, 11(3):229–247.

[91] Resnick, S. and Samorodnitsky, G. (2015). Asymptotic normality of degree counts in a preferential attachment model. *arXiv preprint arXiv:1504.07328*.

[92] Rodriguez, M. G., Balduzzi, D., and Schölkopf, B. (2011). Uncovering the temporal dynamics of diffusion networks. *arXiv preprint arXiv:1105.0697*.

[93] Rudas, A., Tóth, B., and Valkó, B. (2007). Random trees and general branching processes. *Random Structures Algorithms*, 31(2):186–202.

[94] Schuh, H.-J. and Barbour, A. (1977). On the asymptotic behavior of branching processes with infinite mean. *Advances in Applied Probability*, 9(4):681–723.

[95] Seneta, E. (1968). On recent theorems concerning the supercritical galton-watson process. *The Annals of Mathematical Statistics*, 39(6):2098–2102.

[96] Sharpnack, J., Rinaldo, A., and Singh, A. (2012). Changepoint detection over graphs with the spectral scan statistic. *arXiv preprint arXiv:1206.0773*.

[97] Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46.

[98] Shiryaev, A. N. (2007). *Optimal stopping rules*, volume 8. Springer Science & Business Media.

[99] Siegmund, D. (1985). *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media.

[100] Smith, W. L. and Wilkinson, W. E. (1969). On branching processes in random environments. *The Annals of Mathematical Statistics*, 40(3):814–827.

[101] Sun, J., Faloutsos, C., Papadimitriou, S., and Yu, P. S. (2007). Graphscope: parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th acm sigkdd international conference on knowledge discovery and data mining*, pages 687–696. ACM.

[102] Szule, J., Kondor, D., Dobos, L., Csabai, I., and Vattay, G. (2014). Lost in the city: Revisiting milgram's experiment in the age of social networks. *PLoS ONE*, 9(11):e111973.

[103] Szymański, J. (1987). On a nonuniform random recursive tree. *North-Holland Mathematics Studies*, 144:297–306.

[104] Ten Thij, M., Ouboter, T., Worm, D., Litvak, N., van den Berg, H., and Bhulai, S. (2014). Modelling of trends in twitter using retweet graph dynamics. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 132–147. Springer.

[105] Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.

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

[107] van der Hofstad, R., Hooghiemstra, G., and Znamenski, D. (2007). Distances in random graphs with finite mean and infinite variance degrees. *Electronic Journal of Probability*, 12:703–766.

[108] Ver Steeg, G., Ghosh, R., and Lerman, K. (2011). What stops social epidemics? In *AAAI '11*.

[109] Watts, D. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771.

[110] Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM.

[111] Weng, L., Menczer, F., and Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific reports*, 3.

[112] Yang, J. and Leskovec, J. (2010). Modeling information diffusion in implicit networks. In *ICDM '10*, pages 599–608.

[113] Yudovina, E., Banerjee, M., and Michailidis, G. (2015). Changepoint inference for erdős–rényi random graphs. In *Stochastic models, statistics and their applications*, pages 197–205. Springer.

[114] Yule, U. G. (1925). A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87.

[115] Zanette, D. H. (2001). Critical behavior of propagation on small-world networks. *Phys. Rev. E*, 64:050901.