# PROBABILISTIC AND GEOMETRIC APPROACHES TO THE ANALYSIS OF NON-STANDARD DATA

Iain Carmichael

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
2019

Approved by:

Shankar Bhamidi

Jan Hannig

J.S. Marron

Andrew Nobel

Quoc Tran-Dinh

## ABSTRACT

Iain Carmichael: Probabilistic and geometric approaches to the analysis of non-standard data
(Under the direction of: Shankar Bhamidi and J.S Marron)

This dissertation explores topics in machine learning, network analysis, and the foundations of statistics using tools from geometry, probability and optimization.

The rise of machine learning has brought powerful new (and old) algorithms for data analysis. Much of classical statistics research is about understanding how statistical algorithms behave depending on various aspects of the data. The first part of this dissertation examines the *support vector machine* classifier (SVM). Leveraging Karush-Kuhn-Tucker conditions we find surprising connections between SVM and several other simple classifiers. We use these connections to explain SVM's behavior in a variety of data scenarios and demonstrate how these insights are directly relevant to the data analyst.

The next part of this dissertation studies networks which evolve over time. We first develop a method to empirically evaluate vertex centrality metrics in an evolving network. We then apply this methodology to investigate the role of *precedent* in the US legal system.

Next, we shift to a probabilistic perspective on temporally evolving networks. We study a general probabilistic model of an evolving network that undergoes an abrupt change in its evolution dynamics. In particular, we examine the effect of such a change on the network's structural properties. We develop mathematical techniques using *continuous time branching processes* to derive quantitative error bounds for functionals of a major class of these models about their large network limits. Using these results, we develop general theory to understand the role of abrupt changes in the evolution dynamics of these models. Based on this theory we derive a consistent, non-parametric *change point detection* estimator.

We conclude with a discussion on foundational topics in statistics, commenting on debates both old and new. First, we examine the *false confidence theorem* which raises questions for data practitioners making inferences based on epistemic uncertainty measures such as Bayesian posterior distributions. Second, we give an overview of the rise of "data science" and what it means for statistics (and vice versa),

touching on topics such as reproducibility, computation, education, communication and statistical theory.

## ACKNOWLEDGEMENTS

I owe an enormous amount of gratitude to my two advisors, Shankar Bhamidi and Steve Marron, for their enthusiasm, support and encouragement over the past five years. I suspect I went down some unorthodox paths during my PhD and they always supported these forays – or at least the worthwhile ones.

I thank Sayan Banerjee and Jan Hannig for their collaboration (and patience) on research projects over the years. I am appreciative of Andrew Nobel and Quoc Tran-Dinh for serving on my committee and always being willing to chat about both technical and career problems in statistics. I am grateful to everyone[1] who helped me create STOR 390 and turn it into a permanent part of the statistics curriculum, particularly Brendan Brown, Shankar Bhamidi, Robin Cunningham, Mario Giacomazzo, Dylan Glotzer, Varun Goel and Marshall Markham. I am grateful to Christine Keat, Alison Kieber and Samantha Radel, and for helping me to navigate UNC so smoothly. I am thankful for all the support I received from other graduates students including, Kelly Bodwin, Eric Friedlander, Jimmy Jin, Meilei Jung, John Palowitch, Zhengling Qi, and James Wilson. I am especially grateful to Jonathan Williams for his (grudging) willingness to engage in hours (at this point days) of discussions and debates about statistics.

Part of what I love about statistics is how one can parachute into another discipline and work closely with talented and passionate collaborators. I thank CourtListener, Michael Lissner, and James Wudel for teaching me more about the law and providing access to a wealth of legal data. I am grateful to Ben Calhoun, Joseph Geradts, Katherine Hoadley, Linnea Olsen, Charles Perou, and Melissa Troester.

for teaching me about breast cancer pathology and genetics. I thank Heather Couture and Marc Niethammer for introducing me to medical imaging. I am appreciative of Ryan Thornburg for teaching me about journalism in the digital era. I thank Matt Barr, Cameron Freer and Ben Vigoda for everything I learned during my internship at Gamalon. I thank Yiming Hu, Yunxiao Liu, and John Palowitch for being awesome teammates for Citadel's data science competitions. I am appreciative of my undergrad-

[1] Many people gave input to the course's development: https://idc9.github.io/stor390/course_info/acknowledgments.html

uate students and their hard work: Kate Cho, Scott Garcia, James Jushchuk, Michael Kim, Ethan Koch and Charles Tang. I sincerely thank Sandeep Sarangi for his timely and continuous support on UNC's computing cluster without which my productivity would have been cut in half.

I am deeply grateful to Walter Cai, Arturo Fernandez, Talia Fiano, Marina Gaeta, Annie Henderson, Josh Jacobson Carson Mosso, Abby Stevens, and Jason Xu for their willingness to council me on my programming, writing, teaching and life problems. I thank Kerstin Frailey, Giles Hooker and Jason Xu for convincing me that studying statistics in graduate school would be a good idea. I am appreciative of Tim Novikoff for his mentorship from high school all the way through graduate school. I thank Ravi Ramakrishna for his infections passion for math and his career advice. Finally, I am forever indebted to Madison Folks and my parents Deborah and Calum Carmichael for their love, support, advice and proof reading over the years.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS AND SYMBOLS

| | |
|---|---|
| CTBP | Continuous time branching processes |
| $\|\|\cdot\|\|$ | Euclidean norm (unless stated otherwise) |
| DAG | Directed acyclic graph |
| FCT | False confidence theorem |
| FLD | Fischer linear discrimination |
| MD | Mean difference classifier |
| MDP | Maximal data piling classifier |
| PCA | Principal components analysis |
| $\mathbb{R}$ | The set of real numbers |
| RCH | Reduced convex hull |
| SVM | Support vector machine |

CHAPTER 1

# Introduction

One of the major changes in statistics over the past couple decades is the upswing in the use of complex algorithms for data analysis. Classical statistical algorithms such as linear regression and principal components analysis (PCA) can be understood through straightforward linear algebra and geometric considerations. Modern statistical algorithms such as neural networks do not have such simple mathematical characterizations. This raises theoretical, computational and practical challenges for data analysts. One of the aims of my research is to uncover some of the simple, mathematical concepts which underly some of these more sophisticated algorithms. To this end, Chapter 3 of this dissertation presents new geometric insights into the *support vector machine* (SVM) classifier.

Another shift in modern data analysis is the rise of non-standard data such as networks, text and images. These kinds of data often demand new theoretical and methodological tools. For example, the explosion of network data – particularly networks which evolve over time– has given rise to many new mathematical models and statistical methods (e.g. community detection and vertex centrality metrics). Due to the complex statistical dependencies in these network models, it is difficult to understand their properties. Another aim of my research is to develop mathematical and statistical tools which can be used to rigorously analyze the properties of statistical network algorithms. To this end,Chapter 5 presents new statistical methodology to empirically evaluate vertex centrality metrics in a dynamic network. Chapter 6 presents new theory for dynamic network models and network change point detection algorithms.

Finally, these recent changes in statistics have ignited new – and reignited old– debates about the foundations of statistics. This includes topics such as reproducibility, the roles of exploratory/predictive/confirmatory analyses and statistical frameworks such as Bayesian, Frequentist and Fiducial inference. We explore several of these topics in Chapters 7 and 8.

Chapters 3, 5, 6, 7, and 8 represent completed work which has either been published or is currently under review. Chapter 2 gives background on linear classification relevant to Chapter 3 and shows how

a few simple ideas can be used to understand a large number of classification algorithms. Chapter 4 gives a brief overview of the statistical methodology that is developed in Chapter 5. Sections 1.6 and 1.7 discuss the software and educational contributions made during the development of this dissertation. The other sections provide a summary of these projects.

### 1.1 Support vector machine

Chapter 3 studies the *support vector machine* classifier (SVM) and provides new mathematical explanations of SVM's behavior on real data as well as demonstrating connections between SVM and a variety of other classifiers (see Figure 3.1.1). This project started because Prof. Marron noticed that occasionally the SVM *normal vector* direction points in the same direction as the *mean difference*[1] (MD) normal vector and asked me to figure out when this happens. I found this observation surprising because the SVM optimization is a quadratic program and in general does not have an analytic solution. I therefore (naively) tried to prove that SVM could *not* give exactly the MD direction, except for perhaps in very special data scenarios. I ended up proving the opposite. For soft-margin SVM on a binary classification dataset, this MD like behavior *always* happens[2] when the two classes are balanced; when the classes are unbalanced, the MD like behavior occurs approximately (see Theorem 3.4.5).

SVM does not have an analytical solution in general so reasoning about how the SVM solution depends on the data is challenging. For example, when tuning soft-margin SVM it is very easy to make SVM classify every data point into one class (see Section 3.1.1). Even worse, when using cross-validation to tune SVM, sometimes the cross-validation error curve looks significantly different from the test error curve (see Figure 3.1.3d). Using the Karush-Kuhn-Tucker conditions we are able to provide rigorous explanations of each of these phenomena. These mathematical explanations provide further insight into a variety of other SVM phenomena, for example, how SVM cross-validation can be sensitive to properties of the data (see Section 3.6) and connections between SVM (both hard and soft margin) and other classifiers such as naive Bayes, Fisher's linear discriminant and the maximal data piling direction (see Figure 3.1.1).

[1] The MD is perhaps the most simple linear classifier for binary classes; compute each class mean and let the normal vector point from one mean to the other. The MD classifier is discussed in more detail in Chapter 2.

[2] For values of the tuning parameter smaller than a thresholding value which is a deterministic function of the dataset see 3.4.3.

This project is joint work with Prof. Marron and is under review for publication at the time of this writing (Carmichael and Marron, 2017).

### 1.2 Network analysis and the law

When the Supreme Court of the United States decides a case, the judges write an opinion. In the US (and many other countries) these court opinions are new law. Like academic articles, these opinions cite one another and create the Supreme Court citation network. A fundamental part of the US legal system is that judges must use previous legal decisions to decide new cases. In other words, judges are bound by *precedent*. Studying the Supreme Court citation network, or the much larger citation network of all US legal opinions, provides insights into how the law evolves over time. In (Carmichael et al., 2017), reproduced in Chapter 5, we use *vertex centrality metrics* to study the influence of precedent in the US Supreme Court and Federal Appellate circuit.

To obtain this legal citation network data[3], we partnered with a legal non-profit, Courtlistener[4] which provides access to about 3 million US legal cases including the text of the opinion, the citation network and a variety of other data.

In (Carmichael et al., 2017), we find new, empirical evidence of the role that precedent plays in the Supreme Court and US Federal appellate court. These findings come from a new methodology we develop to evaluate different vertex centrality metrics in an evolving network. In particular, we make the surprising discovery that an opinion's out-degree[5] (and related metrics) is more predictive of that opinion receiving future citations than the opinion's in-degree. In other words, the length of an opinion's bibliography, from a network standpoint, is a better measure of a case's popularity than the number of citations that case has already received. This fact is surprising due to the popularity of mathematical network models such as *preferential attachment* (Barabási and Albert, 1999) where the current number of citations plays a fundamental role in acquiring future citations. From a legal scholarship standpoint,

[3]At the time of this writing, US court opinions are expensive to get in bulk; in general one has to pay 10 cents an opinion. This creates difficulties for public defenders, anyone who can't afford a major law firm and academics such as yours truly. This state of affairs is under active debate and litigation e.g. https://www.nytimes.com/2019/02/07/opinion/pacer-court-records.html.

[4]https://www.courtlistener.com/

[5]The citation network is directed with edges going from the citing case to the cited case.

however, this finding is interesting because it speaks to the importance of *precedent* in the Supreme Court (and broader legal system).

The methodology we develop to empirically evaluate vertex centrality metrics, which we call the "sort experiment", is of independent interest for network science and information retrieval. Because the focus of (Carmichael et al., 2017) is on the legal application, Chapter 4 gives a brief overview of this methodology for a statistics audience and discusses some future directions.

This project is the result of a collaboration with James Wudel, James Jushchuk and Michael Kim and has been published (Carmichael et al., 2017).

### 1.3 Dynamic network models and change point detection

Recent years have seen a burgeoning supply of and demand for dynamic networks models[6]. An important question for systems which evolve over time is the ability to detect *change points* (times at which the evolution of the system abruptly changes).

In (Banerjee et al., 2018), reproduced in Chapter 6, we study a broad class of probabilistic models of an evolving network. This *nonuniform random recursive trees* (NRRT) model is parametrized by a function $f : \mathbb{Z}_+ \to \mathbb{R}_+$. In this model, a random tree is grown by successively adding new vertices. When a new vertex, $n$, is added to the system it randomly selects an existing vertex, $v$, to attach to with probability proportional to $f(\text{out-degree}(v))$ where out-degree$(v)$ is the number of nodes[7] currently attached to $v$. A special case of this model is the famous *preferential attachment* model where $f$ is an affine function[8]. In particular, we study a generalization of the NNRT where there is a change point i.e. the network evolves according to one function, $f$, for some time then switches dynamics and evolves according to another function, $g$.

There are two main contributions in (Banerjee et al., 2018). First, we develop mathematical techniques based on *continuous time branching processes* to derive quantitative error bounds for functionals of a major class of these models about their large network limits. Second, we develop general theory to understand the role of abrupt changes in the evolution dynamics of these models. This theory is used to

[6] E.g. networks which evolve over time.

[7] Here edges go from existing vertices to new vertices

[8] Note that a NRRT is not necessarily preferential e.g. $f(x) = \frac{1}{x+1}$.

develop a non-parametric change point detection estimator which requires no knowledge of the attachment functions, pre- or post- change point.

In the context of the second aim, for fixed final network size $n$ and a change point $\tau(n) < n$, we consider models of growing networks. These networks evolve via new vertices attaching to the pre-existing network according to one attachment function $f$ till the system grows to size $\tau(n)$ when new vertices switch their behavior to a different function $g$, until the system reaches size $n$. With general non–explosivity assumptions on the attachment functions $f, g$, we consider both the standard model where $\tau(n) = \Theta(n)$ as well as the *quick big-bang model* when $\tau(n) = n^\gamma$ for some $0 < \gamma < 1$. Proofs rely on a careful analysis of an associated *inhomogeneous* continuous time branching process. Techniques developed in the paper are robust enough to understand the behavior of these models for any sequence of change points $\tau(n) \to \infty$. The paper derives rates of convergence for functionals such as the degree distribution. The same proof techniques should enable one to analyze more complicated functionals such as the associated fringe distributions.

Surprisingly, for the standard model we find that no matter the value of $\gamma$, major structural properties of the network (e.g. the tail behavior of the degree distribution) are determined by the initial function $f$. To probe this *long range dependence* phenomenon we also study the *quick big-bang* model and see that on this time scale these structural properties are no longer affected by the initial function.

The project is a joint work with Prof. Shankar Bhamidi and Prof. Sayan Banerjee and is under review at the time of writing (Banerjee et al., 2018).

### 1.4 False confidence

The foundations of statistics are still up for debate. In addition to the Bayesian and Frequentist paradigms a number of new (and old) statistical frameworks have entered the fray including: generalized fiducial inference (Hannig et al., 2016), Dempster-Schaffer theory (Shafer, 1976), and inferential models (Martin and Liu, 2013).

A recent paper presents the "false confidence theorem" (FCT) which has potentially broad implications for statistical inference using Bayesian posterior uncertainty (Balch et al., 2017). This theorem says that with arbitrarily large (sampling/frequentist) probability, there exists a set which does *not* contain the true parameter value, but which has arbitrarily large posterior probability. Since the use of Bayesian

methods has become increasingly popular in applications of science, engineering, and business, it is critically important to understand when Bayesian procedures lead to problematic statistical inferences or interpretations. In this paper, we consider a number of examples demonstrating the paradoxical nature of false confidence to begin to understand the contexts in which the FCT does (and does not) play a meaningful role in statistical inference. Our examples illustrate that models involving marginalization to non-linear, not one-to-one functions of multiple parameters, play a key role in more extreme manifestations of false confidence.

This project, discussed in Chapter 7, is a joint work with Jonathon Williams and has been published (Carmichael and Williams, 2018).

### 1.5 Data science

Ask a statistician about their feeling towards *data science* and you are likely to provoke a conflicted response. On the one hand, they are likely to tell you that there is not much new to data science – that it's simply a rebranding of statistics. If you push them a little, however, they will likely also admit that "rebranding of statistics" does not quite do justice to the changes in data analysis that have occurred in recent decades. Although a lot has been written about this topic, it can be difficult to cut through the noise and summarize the important points about what this new data science movement means for statistics (and vice versa). For the inaugural issue of the Japanese Journal of Statistics and Data Science, Prof. Marron and I were asked to write a review paper of statistics and data science (Carmichael and Marron, 2018). We touch on a number of topics including: reproducibility, computation, education, communication and statistical theory. The paper is reproduced in Chapter 8.

### 1.6 Statistical software

A lot of modern statistics involves writing software. In an effort to make my software useful for the broader community I have released a number of open source python and R packages. Most of these were methods I needed for my research, but did not exist as software (or exist in the way I needed). While these are not published in an academic journal, I consider these packages to be important contributions of my thesis.

- **jive**: a Python package for dimensionality reduction for multiple data matrices (implements AJIVE). Code: https://github.com/idc9/py_jive.
- **ajive**: an R package implementing AJIVE. Code: https://github.com/idc9/r_jive.
- **what the cluster**: a python package implementing algorithms to determine the “optimal" number of clusters (e.g. gap statistic). Code: https://github.com/idc9/what_the_cluster.
- **jackstraw**: a python package implementing jackstraw type methods which perform statistical tests for dimensionality reduction and other unsupervised algorithms. Code: https://github.com/idc9/jackstraw.
- **diproperm**: a python package implementing DiProPerm for high dimensional hypothesis testing with linear classifiers. Code: https://github.com/idc9/diproperm.

Code for all of my publications can be found at https://github.com/idc9.

### 1.7 Teaching and undergraduate mentorship

With funding from the Data@Carolina initiative, I created and taught the first data science course offered by UNC’s statistics department. My goal was to create a course which teaches “core"[9] data science computational skills for undergraduates. The skills include: basic programming (e.g. visualization, data subsetting/transformation[10], web scraping, regular expressions), statistical modeling (e.g. regression, classification, and clustering), communication (e.g. writing, code clarity, Shiny, R Markdown) as well as a variety of special topics (e.g. natural language processing, data ethics).

All of the course material can be found online at https://idc9.github.io/stor390/. This course is now a permanent part of that curriculum and the material is used by professors at other universities such as Cornell University and Florida Atlanta University. While this course was built for statistics majors, my hope is that similar courses can be developed for other departments such as journalism, psychology, and biology.

During my PhD I mentored 8 undergraduate students. In collaboration with Prof. Bhamidi, I supervised senior theses/independent research projects on networks, text analysis and deep learning for:

[9] What these “core" skills are is very much up for debate.

[10] http://www.mimno.org/articles/carpentry/

Ethan Koch, James Jushchuk, Kate Cho, Scott Garcia, and Michael Kim. I helped Colman Breen and Jenny Chen develop a Shiny app to explore spurious correlations for UNC's science expo.[11] Jointly with Prof. Marc Niethammer I supervised Charles Tang to develop multi-view data analysis software.

[11] https://github.com/idc9/UNC_science_expo_2018

CHAPTER 2

# Linear classifiers

Standard linear classifiers (Friedman et al., 2001) include: *mean difference* (MD), *naive bayes* (NB)(Domingos and Pazzani, 1997; Bickel and Levina, 2004), and Fisher's linear discriminant (FLD). There a several perspectives one can take on a linear classifier including: MLE of a probability distribution, one dimensional projection of the data, optimization and data transformation. In this section we first present a few different perspectives on and connections between these standard classifiers. We then then discuss the maximal data piling direction (MDP) and how the MDP can be viewed as a generalization of FLD in high dimensions.

This Chapter considers binary classification. Suppose we have $n$ labeled data points $\{(\mathbf{x}_i, y_i)_{i=1}^n\}$ and index sets $I_+, I_-$ such that $y_i = 1$ if $i \in I_+$, $y_i = -1$ if $i \in I_-$ and $\mathbf{x}_i \in \mathbb{R}^d$. Let $n_+ = |I_+|$ and $n_- = |I_-|$ be the class sizes (note that $n_- + n_+ = n$). Let $X \in \mathbb{R}^{n \times d}$ be the data matrix with the observations on the rows. Let $\bar{\mathbf{x}}_+ := \frac{1}{n_+} \sum_{i \in I_+} \mathbf{x}_i \in \mathbb{R}^d$ be the mean of the data points in the positive class (similarly for $\bar{\mathbf{x}}_-$).

A *linear classifier* is defined through a hyperplane with *normal vector* $\mathbf{w} \in \mathbb{R}^d$ and *intercept* $b \in \mathbb{R}$; it classifiers all points on one side of the hyperplane to one class and all points on the other side to the other class. More formally, a linear classifier's *decision function* is given by $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ with classification rule sign $\hat{y} = (f(\mathbf{x}))$.

Given two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ we consider their *directions to be equivalent* if there exists $a \in \mathbb{R}, a \neq 0$ such that $a\mathbf{w} = \mathbf{v}$ (and we will write $\mathbf{w} \propto \mathbf{v}$). Using this equivalence relation we can quotient $\mathbb{R}^d$ into the space of directions (formally real projective space). Intuitively, this is the space of lines through the origin. Another way to think about the direction of a linear classifier is to assume the normal vector has unit norm (and pick an orientation).

One key idea in this section is to compare *directions* of linear classifiers' normal vectors. Note that two classifiers may have the same direction, but lead to different classification algorithms (i.e. the intercepts may differ).

## 2.1 Mean Difference

The mean difference classifier selects the direction pointing between the two class means.

**Definition 2.1.1.** *The mean difference direction is given by*

$$\mathbf{w}_{MD} := \bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_- \tag{2.1.1}$$

The MD classifier is one of the most basic classifiers and can be used to both understand and motivate more sophisticated classifiers. MD can be understood from multiple perspectives.

If the data generative distribution is two independent multivariate Gaussians with identity covariance but (possibly) different means then $\mathbf{w}_{MD}$ is the decision rule corresponding to the maximum likelihood estimates of the distributions. The intercept can be chosen in a number of different ways e.g. maximizing a likelihood or minimizing the cross-validation misclassification error. Another option is to select the intercept so that the separating hyperplane goes half way between the two class means. In the latter case MD will classify a new data point to the class with the nearest mean and is hence sometimes called the *nearest centroid* classifier.

Often it is worth viewing a linear classifier as projecting the data onto a one dimensional subspace. We might seek a classifier that separates the two classes as much as possible. One way to make this heuristic concrete is to find the direction that separates the projected class means as much as possible leading to the following optimization problem:

$$\begin{aligned} \underset{\mathbf{w}\in\mathbb{R}^d}{\text{maximize}} \quad & \mathbf{w}^T\bar{\mathbf{x}}_+ - \mathbf{w}^T\bar{\mathbf{x}}_- \\ \text{subject to} \quad & ||\mathbf{w}|| = 1. \end{aligned} \tag{2.1.2}$$

It's straightforward to show that the solution to this problem is given by $\mathbf{w}_{MD}$.

## 2.2 Naive Bayes, linear discriminant analysis, and maximal data piling

Several of the following classifiers (NB, FLD) can be viewed as applying the MD classifier after a data transformation or using a particular *Mahalanobis distance*. Some of these classifiers (FLD, MDP) can

also be motivated by considering the one dimensional projections and writing down a modified version of optimization Problem (2.1.2).

The three classifiers (NB, FLD, MDP) are all in the form of a precision matrix estimate (inverse covariance) times the MD direction i.e.

$$\begin{aligned} \mathbf{w} &= \Sigma^{-1}(\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-) \\ &= P\mathbf{w}_{MD}. \end{aligned} \tag{2.2.1}$$

Where $\Sigma \in \mathbb{R}^{d\times d}$ is some type of covariance (global, pooled, diagonal, etc) matrix estimate and $P$ is some flavor of inverse (e.g. Moore-Penrose) of this covariance matrix.

In this section we assume the data points are in *general position* (e.g. they were generated from a continuous probability distribution). This means[1] $\dim(\text{span}(\{\mathbf{x}_i\}_1^n)) = \min(n,d)$. When a covariance matrix in this section is not invertible we can switch to a generalized inverse to still get a valid classifier.

We now define several covariance matrix estimates. The global version is the standard covariance matrix estimate of a set of n observations.

**Definition 2.2.1.** *The global covariance matrix estimate is given by*

$$\begin{aligned} \widehat{\Sigma}_{global} &:= \frac{1}{n-1}(X-\bar{X})^T(X-\bar{X}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T. \end{aligned} \tag{2.2.2}$$

Note that $\text{rank}(\widehat{\Sigma}_{global}) = \dim(\text{span}(\{\mathbf{x}_i - \bar{\mathbf{x}}\}_1^n))$. If $n > d$ and the data are in general position then $\widehat{\Sigma}_{global}$ is full rank. If $n \le d$ then $\text{rank}(\widehat{\Sigma}_{global}) = n-1$.

The pooled version estimates individual covariance matrices for each class and then takes a weighted average.

**Definition 2.2.2.** *The pooled covariance matrix estimate for two classes of data is given by*

$$\begin{aligned} \widehat{\Sigma}_{pool} &:= \left(\frac{n_+-1}{n-2}\right)\Sigma_+ + \left(\frac{n_--1}{n-2}\right)\Sigma_- \\ &:= \frac{1}{n-2}\left[(X_+-\overline{X}_+)^T(X_+-\overline{X}_+) + (X_--\overline{X}_-)^T(X_--\overline{X}_-)\right] \\ &= \frac{1}{n-2}\left[\sum_{i\in I_+}(\mathbf{x}_i-\bar{\mathbf{x}}_+)(\mathbf{x}_i-\bar{\mathbf{x}}_+)^T + \sum_{i\in I_-}(\mathbf{x}_i-\bar{\mathbf{x}}_-)(\mathbf{x}_i-\bar{\mathbf{x}}_-)^T\right]. \end{aligned} \tag{2.2.3}$$

[1] If we work with the globally centered data vectors we lose a dimension i.e. $\dim(\text{span}(\{\mathbf{x}_i - \text{Var}\mathbf{x}\}_1^n)) = \min(n,d)-1$

In some cases we only want to estimate a diagonal covariance matrix i.e. we individually estimate the variance of each variable. For a moment consider two classes of points in one dimension i.e. $x_1, \ldots, x_n \in \mathbb{R}$. The global variance estimate is the familiar

$$\widehat{\sigma}^2_{global} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Similarly the pooled variance estimate is

$$\widehat{\sigma}^2_{pool} := \frac{1}{n-2} \left( \sum_{i \in I_+}^{n} (x_i - \bar{x}_+)^2 + \sum_{i \in I_+}^{n} (x_i - \bar{x}_+)^2 \right).$$

Returning to the $d$ dimensional case, let

$$\widehat{\sigma}^2_{global,j} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)^2$$

be the global variance estimate of the $jth$ variable (similarly for $\widehat{\sigma}^2_{pool,j}$). We now define two diagonal covariance matrix estimates.

**Definition 2.2.3.** *The global diagonal covariance matrix estimate for two classes of data is given by*

$$\widehat{\Sigma}_{global,diag} := diag\left[\widehat{\sigma}^2_{global,j}\right]_{j=1,\ldots,d}. \tag{2.2.4}$$

**Definition 2.2.4.** *The pooled diagonal covariance matrix estimate for two classes of data is given by*

$$\widehat{\Sigma}_{pool,diag} := diag\left[\widehat{\sigma}^2_{pool,j}\right]_{j=1,\ldots,d}. \tag{2.2.5}$$

Now we can define corresponding linear classifiers for the above covariance matrix estimates. The naive Bayes classifier is defined using the diagonal, pooled covariance matrix.

**Definition 2.2.5.** *The naive bayes direction is given by*

$$\mathbf{w}_{NB} := \widehat{\Sigma}^{-1}_{pool,diag} \mathbf{w}_{MD}. \tag{2.2.6}$$

We define *Fisher's linear discriminant* using the full, pooled covariance matrix estimate.

**Definition 2.2.6.** *When $d < n$ Fisher's linear discriminant direction is given by*

$$\mathbf{w}_{FLD} := \widehat{\Sigma}_{pool}^{-1}\mathbf{w}_{MD}. \tag{2.2.7}$$

If $\widehat{\Sigma}_{pool}$ is not invertible we switch to a generalized inverse (more on this below).

Now for all $n, d$ we define the maximal data piling direction (Ahn and Marron, 2010a) using the global covariance matrix.

**Definition 2.2.7.** *The maximal data piling direction is given by*

$$\mathbf{w}_{MDP} := \widehat{\Sigma}_{global}^{-}\mathbf{w}_{MD}, \tag{2.2.8}$$

where $\widehat{\Sigma}_{global}^{-}$ is the generalized Moore-Penrose inverse of $\widehat{\Sigma}_{global}$. While this definition of $\mathbf{w}_{MDP}$ is well defined for all $n, d$ the classifier has two regimes of behavior depending on whether $n > d$(discussed below).

NB and FLD have statistical interpretations as MLEs of Gaussian distributions. FLD assumes the data come from two classes with different means but identical covariance matrices. NB is similar but has the more restrictive assumption that the covariance matrix is diagonal. These statistical assumptions are by no means required to be true to use these classifiers.

### 2.3 Data transformation perspective: covariance matrix + mean difference

It is common to transform the data before fitting a statistical model. We can see that both NB and FLD are equivalent to MD after applying a linear transformation to the data. FLD applies the inverse square root (defined above) of $\widehat{\Sigma}_{pool}$ to the data. If we center the data by the global mean then is the so called "whitening transformation" which makes the covariance matrix of the transformed data equal to the identity. NB applies the inverse square root of $\widehat{\Sigma}_{pool,diag}$ which scales each variable by it's pooled covariance estimate. An equivalent perspective is: FLD is MD using the Mahalanobis distance defined by $\widehat{\Sigma}_{pool}$ (similarly for NB and $\widehat{\Sigma}_{pool,diag}$). In this section we assume the data have been first centered by the global mean $\bar{\mathbf{x}}$ (this includes test data points).

Consider applying a linear transformation, $S \in \mathbb{R}^{d\times d}$, to the data[2] i.e. let $\tilde{\mathbf{x}}_i = S\mathbf{x}_i$. Next compute the mean difference of the transformed data i.e.

$$\tilde{\mathbf{w}}_{MD} = \tilde{\mathbf{x}}_+ - \tilde{\mathbf{x}}_- = S\bar{\mathbf{x}}_+ - S\bar{\mathbf{x}}_- = S\mathbf{w}_{MD}.$$

Next consider a test point $\mathbf{x} \in \mathbb{R}^d$. Note that since we have applied a data transformation to the training data we apply the same transformation to $\mathbf{x}$ i.e. $\tilde{\mathbf{x}} = S\mathbf{x}$. Computing the prediction (we ignore the intercept term)

$$\tilde{\mathbf{x}}^T\tilde{\mathbf{w}}_{MD} = (S\mathbf{x})^T(S\mathbf{w}_{MD}) = \mathbf{x}(S^TS\mathbf{w}_{MD}). \tag{2.3.1}$$

Now suppose $S$ is the inverse square root of some matrix $\Sigma$ i.e. $S^TS = \Sigma^{-1}$. Then from (2.3.1) we get

$$\tilde{\mathbf{x}}^T\tilde{\mathbf{w}}_{MD} = \Sigma^{-1}\mathbf{w}_{MD} \tag{2.3.2}$$

**Remark 2.3.1.** Equation (2.3.2) shows that applying an inverse square root matrix transformation $\Sigma^{-\frac{1}{2}}$ then fitting the mean difference classifier is equivalent to fitting a classifier of the form $\Sigma^{-1}\mathbf{w}_{MD}$.

The upshot of this perspective is: we understand many standard classifiers by first applying a data transformation then computing the mean difference. So why does this perspective matter? Here are a few reasons.

- We can create and understand many more complicated classifiers by combining two very simple ideas: data transformation and the mean difference.
- This perspective emphasizes the geometry of the problem. The MD classifier works best when the data are very nice (i.e. two spheres). Depending on the shape of the data we apply a transformation that best spheres the data.
- We may choose non-standard estimates for the centers and covariance structure of the data and apply these estimates analogously. For example, we might want robust estimates of the class means (e.g. *spatial median* Brown 1983) or covariance matrices (e.g. *minimum volume ellipsoid* Van Aelst and Rousseeuw 2009).

[2]Note that when computing the prediction of a test point we apply the same transformation to the test point.

- MD can be kernelized to allow for efficient computation of non-linear data transformations. This observation is not strictly within this framework, but is worth mentioning.

### 2.4 Linear discriminant vs. maximal data piling

In the low dimensional setting the FLD direction solves an optimization problem that comes from thinking about the one dimensional projection of the data (Bishop, 2006). Recall that the mean difference is the direction that gives the larges separation between the two class means. One issue with the mean difference is there can be a lot of overlap between the two projected classes. We might therefore seek a classifier that attempts to both separate the two class means while keeping projections within the same class close together.

For a given direction $\mathbf{w}$, let the "within-class variance" be $s_+^2 := \sum_{i \in I_+} (\mathbf{x}^T\mathbf{w} - \bar{\mathbf{x}}_+^T\mathbf{w})^2$ the variance of the projected points in the positive class (similarly for $s_-^2$). Call the squared distance between the projected means the "between-class" variance (e.g. $(\bar{\mathbf{x}}_+^T\mathbf{w} - \bar{\mathbf{x}}_-^T\mathbf{w})^2$). The Fisher criterion is then the ratio:

$$\begin{aligned} J(\mathbf{w}) &= \frac{\text{between-class variance}}{\text{within-class variance}} \\ &= \frac{(\bar{\mathbf{x}}_+^T\mathbf{w} - \bar{\mathbf{x}}_-^T\mathbf{w})^2}{s_-^2 + s_+^2} \\ &= \frac{\mathbf{w}^T\mathbf{w}_{MD}\mathbf{w}_{MD}{}^T\mathbf{w}}{\mathbf{w}^T\widehat{\Sigma}_{pool}\mathbf{w}^T}. \end{aligned} \tag{2.4.1}$$

Where the last line can be seen with a little algebra. We again restrict $\mathbf{w}$ to have unit norm leading to the following optimization problem of maximizing $J(\mathbf{w})$:

$$\begin{aligned} &\underset{\mathbf{w} \in \mathbb{R}^d}{\text{maximize}} \quad J(\mathbf{w}) \\ &\text{subject to} \quad ||\mathbf{w}|| = 1. \end{aligned} \tag{2.4.2}$$

(Bishop, 2006) shows the solution to Problem 2.4.2 can been seen to be $\mathbf{w}_{FLD}$ given in Definition 2.2.6.

In high dimensions, when $d > n$, there is immediately a problem: the matrix $\widehat{\Sigma}_{pool}$ is no longer invertible and there are directions where the denominator of Equation 2.4.1 is zero so objective function $J(\mathbf{w})$ is infinite! Geometrically each class is being projected to a single point so the within class variance is zero, so called *data piling*.

*Data piling*, first discussed by (Marron et al., 2007), is when multiple points have the same projection on the line spanned by the normal vector. For example, all points on SVM's margin have the same image under the projection map. (Ahn and Marron, 2010a) showed that when $d \geq n-1$ there are directions such that each class is projected to a single point i.e. there is *complete data piling*.

**Definition 2.4.1.** *A vector* $\mathbf{w} \in \mathbb{R}^d$ *gives complete data piling for two classes of data if there exist* $a, b \in \mathbb{R}$, *with* $a \neq 0$ *such that*

$$\mathbf{w}^T \mathbf{x}_i = a y_i + b \text{ for each } i = 1, \ldots, n,$$

*where* $b$ *is the midpoint of the projected classes and* $a$ *is half the distance between the projected classes.*

When the data matrix $X$ has full column span there exist directions of complete data piling (Ahn and Marron, 2010a). This can be seen by considering the linear system $X\mathbf{w} = y + \mathbf{1}_n b$ with variables $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$.

The existence of complete data piling directions ruins the FLD criterion. We can make the following modification to the problem which will motivate the maximal data piling classifier. Instead of considering the ratio in Equation 2.4.1 let's restrict ourselves to directions of complete data piling (Definition 3.2.2). Of all complete data piling directions let's select the one that gives maximal separation between the two class means.

When there are directions of maximal data piling we consider the following optimization problem motivated by the above discussion.

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^d, a, b \in \mathbb{R}}{\text{maximize}} \quad & \mathbf{w}^T (\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-) \\ \text{subject to} \quad & \mathbf{w}^T \mathbf{x}_i = a y_i + b \text{ for each } i = 1, \ldots, n. \end{aligned} \tag{2.4.3}$$

(Ahn and Marron, 2010a) showed that Problem 2.4.3 has a unique solution given by

$$\mathbf{w}_{MDP} = \widehat{\Sigma}^-_{global} \mathbf{w}_{MD}$$

where $\widehat{\Sigma}^-_{global}$ is the Moore-Penrose inverse of the global covariance matrix. Properties of this classifier have been studied in a number of papers such as (Ahn et al., 2012; Lee et al., 2013; Ahn and Marron, 2010a).

Problem 2.4.3 only makes sense when directions of complete data piling exist. However the direction $\widehat{\Sigma}^{-}_{global}\mathbf{w}_{MD}$ does make sense in low dimensional settings when the generalized inverse is equal to the inverse. In other words when in low dimensions we can define $\mathbf{w}_{MDP} = \widehat{\Sigma}^{-1}_{global}\mathbf{w}_{MD}$. It's worth contrasting this direction with the FLD direction $\mathbf{w}_{FLD} = \widehat{\Sigma}^{-1}_{pool}\mathbf{w}_{MD}$ which uses the pooled covariance matrix.

This raises a natural question: how is the MDP direction related to the FLD direction in the low dimensional setting? Surprisingly, they are the same.

**Theorem 2.4.2.** *When $d \leq n-1$ the MDP direction and FLD directions are the same i.e.*

$$\widehat{\Sigma}^{-1}_{global}(\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-) \propto \widehat{\Sigma}^{-1}_{pool}(\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-) \tag{2.4.4}$$

In other words, transforming the mean difference direction $\mathbf{w}_{MD}$ by the global covariance matrix is equivalent to transforming $\mathbf{w}_{MD}$ by the pooled sample covariance matrix. To prove this we need two lemmas (a similar version of this proof originally appeared in Ahn and Marron 2010a).

**Lemma 2.4.3.**

$$\widehat{\Sigma}_{global} = \frac{n-2}{n-1}\widehat{\Sigma}_{pool} + \frac{n_+ n_-}{n(n-1)}\mathbf{w}_{MD}\mathbf{w}_{MD}{}^T. \tag{2.4.5}$$

*Proof.* Note that

$$\widehat{\Sigma}_{global} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$\widehat{\Sigma}_{global} = \frac{1}{n-1}\left(\sum_{i\in I_+}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T + \sum_{i\in I_-}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T\right)$$

Consider the first term on the right. We first add zero to get

$$\sum_{i\in I_+}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \sum_{i\in I_+}\left[(\mathbf{x}_i - \bar{\mathbf{x}}_+) - (\bar{\mathbf{x}} - \bar{\mathbf{x}}_+)\right]\left[(\mathbf{x}_i - \bar{\mathbf{x}}_+) - (\bar{\mathbf{x}} - \bar{\mathbf{x}}_+)\right]^T.$$

Expanding this term with the identity $\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{a})(\mathbf{x}_i - \mathbf{a})^T = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T - 2\mathbf{a}\sum_{i=1}^{n}\mathbf{x}_i^T + n\mathbf{a}\mathbf{a}^T$ we get

$$= \sum_{i\in I_+}(\mathbf{x}_i - \bar{\mathbf{x}}_+)(\mathbf{x}_i - \bar{\mathbf{x}}_+)^T - 2(\bar{\mathbf{x}} - \bar{\mathbf{x}}_+)\sum_{i\in I_+}(\mathbf{x}_i - \bar{\mathbf{x}}_+)^T + n_+(\bar{\mathbf{x}} - \bar{\mathbf{x}}_+)(\bar{\mathbf{x}} - \bar{\mathbf{x}}_+)^T$$

The second term cancels ($\sum_{i\in I_+}(\mathbf{x}_i - \bar{\mathbf{x}}_+) = 0$). With some straightforward algebra we see that $\bar{\mathbf{x}} - \bar{\mathbf{x}}_+ - \frac{n_-}{n}(\bar{\mathbf{x}}_- - \bar{\mathbf{x}}_+)$. So we are left with

$$\sum_{i\in I_+}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \sum_{i\in I_+}(\mathbf{x}_i - \bar{\mathbf{x}}_+)(\mathbf{x}_i - \bar{\mathbf{x}}_+)^T + \frac{n_-^2 n_+}{n^2}\mathbf{w}_{MD}\mathbf{w}_{MD}{}^T.$$

We can do the analogous calculation for the negative class and put this together to get

$$\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \sum_{i\in I_+}(\mathbf{x}_i - \bar{\mathbf{x}}_+)(\mathbf{x}_i - \bar{\mathbf{x}}_+)^T + \sum_{i\in I_-}(\mathbf{x}_i - \bar{\mathbf{x}}_-)(\mathbf{x}_i - \bar{\mathbf{x}}_-)^T + \frac{n_-^2 n_+ + n_+^2 n_-}{n^2}\mathbf{w}_{MD}\mathbf{w}_{MD}{}^T.$$

Note that the first two terms are the pooled covariance matrix (up to a constant). A little algebra shows $\frac{n_-^2 n_+ + n_+^2 n_-}{n^2} = \frac{n_+ n_-}{n}$. Applying these observations allows us to conclude the result. ■

Next we need a matrix identity (a slight variation of exercise 5.16 in Searle 1982).

**Lemma 2.4.4.** *Suppose $B \in \mathbb{R}^{d\times d}$ is invertible, $A \in \mathbb{R}^{d\times k}$ and $c \in \mathbb{R}$ then*

$$(B + cAA^T)^{-1}A = B^{-1}A(I + cA^TB^{-1}A)^{-1}. \tag{2.4.6}$$

*Proof.*

$$\begin{aligned}(B + cAA^T)B^{-1}A(I + cA^TB^{-1}A)^{-1} &= (I + cAA^TB^{-1})A(I + cA^TB^{-1}A)^{-1}\\ &= (A + cAA^TB^{-1}A)(I + cA^TB^{-1}A)^{-1}\\ &= A(I + cA^TB^{-1}A)(I + cA^TB^{-1}A)^{-1}\\ &= A\end{aligned}$$

■

We can now prove Theorem 2.4.2

*Proof.*

$$\begin{aligned}
\widehat{\Sigma}_{global}^{-1}\mathbf{w}_{MD} &= \left(\frac{n-2}{n-1}\widehat{\Sigma}_{pool} + \frac{n_+ n_-}{n(n-1)}\mathbf{w}_{MD}\mathbf{w}_{MD}{}^T\right)^{-1}\mathbf{w}_{MD} \\
&= \frac{n-1}{n-2}\widehat{\Sigma}_{pool}^{-1}\mathbf{w}_{MD}\left(\frac{1}{\frac{n_+ n_-}{n(n-1)}\mathbf{w}_{MD}{}^T\widehat{\Sigma}_{pool}\mathbf{w}_{MD}}\right) \\
&\propto \widehat{\Sigma}_{pool}^{-1}\mathbf{w}_{MD}
\end{aligned}$$

The first equality follows from Fact 2.4.3 and the second equality follows from Fact 2.4.4 with $A = \mathbf{w}_{MD}$.

■

To summarize this section we now have two views on how MDP is related to and/or can be seen as a generalization of FLD.

1. The FLD optimization problem (Problem 2.4.2) is not well defined in high dimensional settings because of complete data piling directions. MDP modifies this problem in a sensible way in high dimensional settings by searching over all complete data piling directions and finding the one that gives maximal class separation.

2. MDP and FLD are equivalent in low dimensional settings (Theorem 2.4.2). In high dimensional settings MDP replaces the inverse with the Moore-Penrose inverse of the $\widehat{\Sigma}_{global}$.

CHAPTER 3

# Geometric Insights into Support Vector Machine Behavior using the KKT Conditions

The *support vector machine* (SVM) is a powerful and widely used classification algorithm. this chapter uses the Karush-Kuhn-Tucker conditions to provide rigorous mathematical proof for new insights into the behavior of SVM. These insights provide perhaps unexpected relationships between SVM and two other linear classifiers: the *mean difference* and the *maximal data piling direction*. For example, we show that in many cases SVM can be viewed as a cropped version of these classifiers. By carefully exploring these connections we show how SVM tuning behavior is affected by characteristics including: balanced vs. unbalanced classes, low vs. high dimension, separable vs. non-separable data. These results provide further insights into tuning SVM via cross-validation by explaining observed pathological behavior and motivating improved cross-validation methodology. Finally, we also provide new results on the geometry of *complete data piling directions* in high dimensional space.

## 3.1 Introduction

The *support vector machine* (SVM) is a popular and well studied classification algorithm (for an overview see Schölkopf and Smola 2002; Shawe-Taylor and Cristianini 2004; Steinwart and Christmann 2008; Mohri et al. 2012; Murphy 2012). Classical classification algorithms, such as *logistic regression* and *Fisher linear discrimination* (FLD) are motivated by fitting a statistical distribution to the data. Hard margin SVM on the other hand is motivated directly as an optimization problem based on the idea that a good classifier should maximize the margin between two classes of separable data. Soft margin SVM balances two competing objectives; maximize the margin while penalizing points on the wrong side of the margin.

Interpretability, explainability, and more broadly understanding why a model makes its decisions are active areas of research in machine learning (Guidotti et al., 2018; Doshi-Velez and Kim, 2017). There is a large body of research providing theoretical guarantees and computational advances for studying

SVM (Vapnik, 2013; Steinwart and Christmann, 2008). Several papers have shed some light on SVM by placing it in a probabilistic framework (Sollich, 2002; Polson et al., 2011; Franc et al., 2011). Here we take a different approach based on optimization and geometry, to understand the inner workings of SVM.

The setting of this chapter is the two class classification problem. We focus on linear classifiers, but the results extend to corresponding kernel classifiers. We consider a wide range of data analytic regimes including: high vs. low dimension, balanced vs. unbalanced class sizes and separable vs. non-separable data.

Using the *KKT conditions*, this chapter demonstrates novel insights into how SVM's behavior is related to a given dataset and furthermore how this behavior is affected by the tuning parameter. We discover a number of connections between SVM and two other classifiers: the *mean difference* (MD) and *maximal data piling classifier* (MDP). These connections are summarized in Figure 3.1.1. In particular, when $C$ is small, soft margin SVM behaves like a (possibly cropped) MD classifier (Theorem 3.4.8). When the data are high dimensional, hard SVM (and soft margin with large $C$) behaves like a cropped MDP classifier (Theorem 3.3.4, Corollary 3.3.6, Theorem 3.4.8). The connection between SVM and the MD further implies connections between SVM, after a data transformation, and a variety of other classifiers such as *naive Bayes* (NB) (see Section 3.2.1). The connection between SVM and MDP provides novel insights into the geometry of the MDP classifier (Sections 3.3.1, 3.7.1). These insights explain several observed, surprising SVM behaviors which motivated this chapter (Section 3.1.1). They furthermore have applications to improving SVM cross-validation methodology and lead us to propose a modified SVM intercept term which can improve test set performance (Section 3.6).

### 3.1.1 Motivating Example

The motivation for this chapter is to understand surprising, observed SVM behavior. This section uses a simple, two dimensional example to demonstrate a number of instances of pathological or surprising SVM behavior which the rest of the paper explains and then builds upon.

Figures 3.1.2 and 3.1.3 show the result of fitting SVM for a range of tuning parameters. The data in both figures are generated from a two dimensional Gaussian with identity covariance such that the class means are 4 apart. In Figure 3.1.2 the classes are balanced (20 points in each class). The data points in

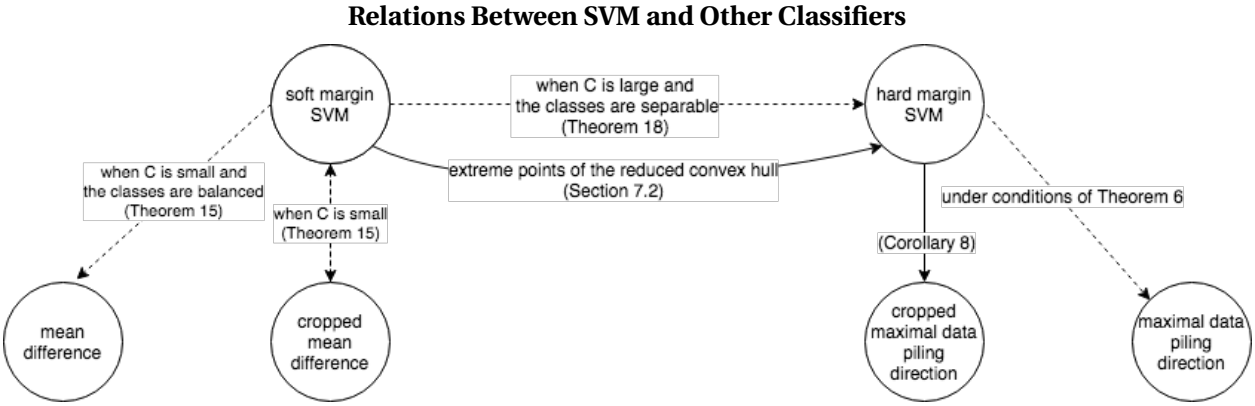


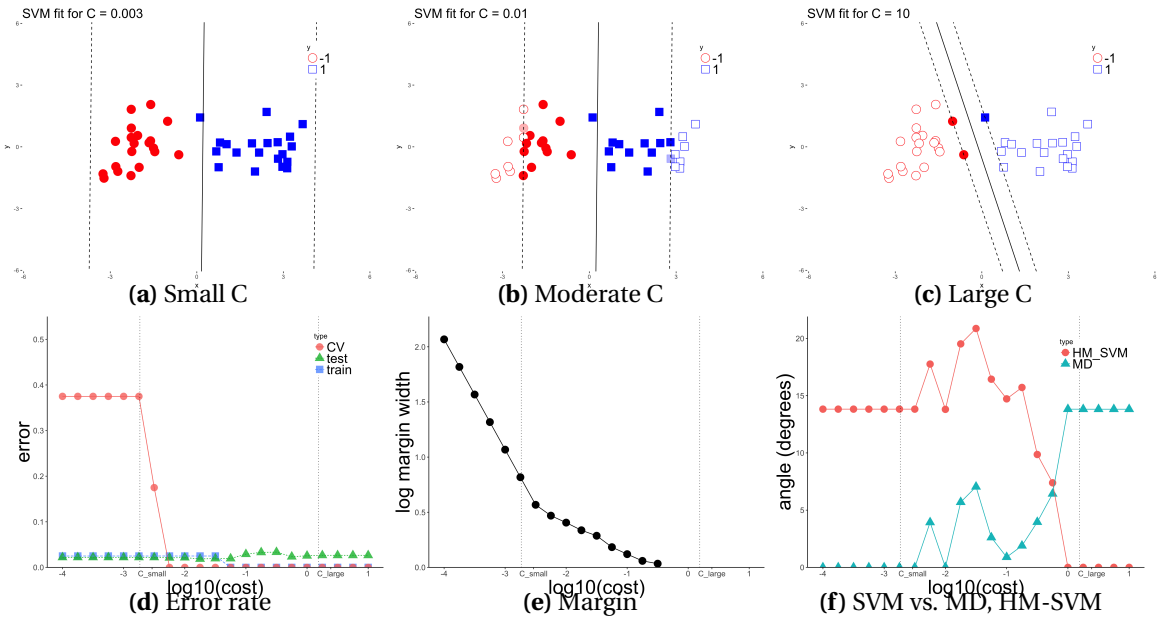

**Figure 3.1.1:** SVM reduces to another classifier under the condition stated in the arrow. Solid line means the relation always holds. Dashed line means the relation may or may not hold depending on the data. For example, SVM reduces to the mean difference when the classes are balanced and $C$ is sufficiently small ($C \leq C_{\text{small}}$) which is shown in Theorem 3.4.5.

Figure 3.1.3 are the same points as the first figure, but one additional point is added to the positive class (blue squares) so the classes are unbalanced. In both cases the classes are linearly separable.

The top row of panels show the data along with the SVM separating hyperplane (solid line) for three different values of $C$. The *marginal hyperplanes* are shown as dashed lines and the filled in symbols are *support vectors*. The bottom three panels show various functions of $C$. The bottom left panel shows three error curves: training, cross-validation (5-folds), and test set error. The bottom middle panel shows the margin width. Finally, the bottom right panel shows the angle between the soft margin SVM direction and both the hard margin SVM direction and the mean difference direction. The vertical dashed lines indicate the values of $C_{small}$ and $C_{large}$ which are discussed below. See references above or Sections 3.3.3, 3.4.1 for definitions of the margin and support vectors.

Important features of these plots include:

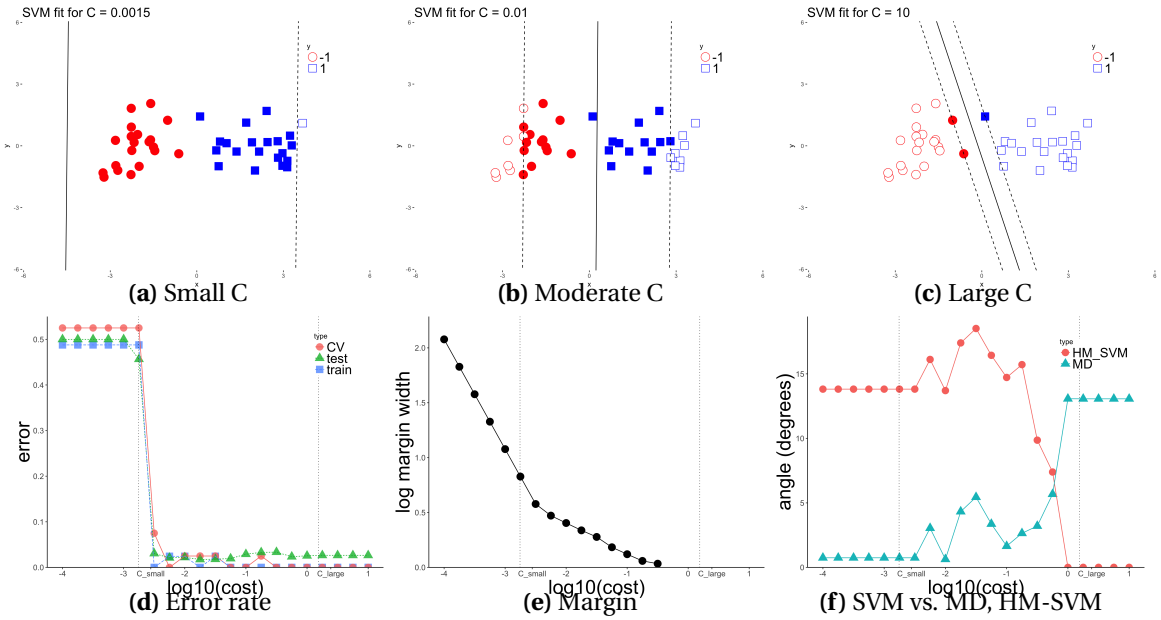
1. For balanced classes (Figure 3.1.2), the training, cross-validation and test set error is low for most values of $C$, then suddenly shoots up to around 50% error for a small enough values of $C$ (see Figure 3.1.2d). For unbalanced classes (Figure 3.1.3), this tuning error explosion for small $C$ only happens for cross validation, not the tuning or test sets (see Figure 3.1.3d). This pathological behavior is concerning for a number of reasons, including it demonstrates an example when performance with cross-validation may not reflect test set performance. Moreover, it is not clear why this behavior is happening.




**(a)** Small C **(b)** Moderate C **(c)** Large C

**(d)** Error rate **(e)** Margin **(f)** SVM vs. MD, HM-SVM

**Figure 3.1.2:** (Balanced classes) The top rows show the SVM fit for various values of $C$. The bottom row shows diagnostics which are described in the text. Figure 3.1.2d shows that the cross-validation error curve can be very different from the training and test error. Figure 3.1.2f shows that for small enough values of $C$, the SVM and MD directions are the same.




**(a)** Small C **(b)** Moderate C **(c)** Large C

**(d)** Error rate **(e)** Margin **(f)** SVM vs. MD, HM-SVM

**Figure 3.1.3:** (Unbalanced classes). The panels are the same as in Figure 3.1.2, but the data now have one additional point added. When $C$ is small, the top left panel shows SVM classifies every point to the larger class (the separating hyperplane is pushed past the smaller class). For this unbalanced example the cross-validation, train and test error all behave similarly, unlike the balanced case (compare Figure 3.1.3d to 3.1.2d). When $C$ is small, the angle between SVM and the MD is small but not exactly zero (compare Figure 3.1.3f to 3.1.2f).

2. Figure 3.1.2f show that the SVM decision boundary can be parallel to the mean difference decision boundary when the data are balanced. This behavior is surprising because the SVM optimization problem is not immediately connected to the means of the two classes. Similarly, Figure 3.1.3f demonstrates an example when the SVM and MD decision boundaries are almost parallel for unbalanced classes.

3. Both Figures 3.1.2f and 3.1.3f show that soft margin SVM becomes exactly equivalent to hard margin SVM for some finite value of $C$ when the data are separable.

Theorem 3.4.5 gives a complete answer to why and when the first two of these behaviors occur. Moreover, it demonstrates that this behavior occurs for every dataset. For the first example, if the data are unbalanced then the intercept term will always go off to infinity for small enough values of the tuning parameter; while SVM finds a good direction, its performance is betrayed by its intercept. For the second example, when $C$ is smaller than a threshold value $C_{\text{small}}$ (Definition 3.4.3), the SVM direction will be exactly equivalent to the MD direction when the data are balanced. Similarly, when the data are unbalanced and $C < C_{\text{small}}$ the SVM direction is close to the MD direction. In this latter case, Equations 3.4.3, 3.4.4 show the SVM direction must satisfy constraints that make it a cropped mean difference direction.

A formula for this threshold $C_{\text{small}}$ governing when SVM behaves like the MD is given in Definition 3.4.3 as a function of the diameter of the training data. Similarly, a formula for a threshold $C_{\text{large}}$ governing when soft margin SVM becomes hard margin SVM is given in Definition 3.4.4 as a function of the gap between the two training classes. These two thresholding values are shown as dotted vertical lines in the bottom three panels of Figures 3.1.2 and 3.1.3.

Careful study of these behaviors, including the given formulas for the two thresholds, shows ways in which soft margin SVM's behavior can change depending on characteristics of the data including: balanced vs. unbalanced classes, whether $d \geq n-1$, the two class diameter, whether the classes are separable and the gap between the two classes when they are separable. These results then lead to new insights into SVM tuning (Section 3.6).

### 3.1.2 Related Literature

(Hastie et al., 2004) show how to efficiently compute the entire SVM tuning path. While a consequence of their technical results shows that for small enough $C$, SVM behaves like the MD, they don't

make the explicit connection to the MD classifier. For balanced classes they prove SVM is equivalent to the MD. For unbalanced we give a stronger, more specific characterization as a cropped MD (see Theorem 3.4.5 and Lemma 3.4.6). Additionally, they did not find the important, general threshold values $C_{\text{small}}$ or $C_{\text{large}}$ which depend on the diameter (gap) of the data which have useful consequences for cross-validation.

Connections between SVM and other classifiers have been studied before, for example, (Jaggi, 2014) studies connections between SVM and logistic regression with an L1 penalty.

We thank the reviewers for pointing us to the nu-SVM literature (Schölkopf et al., 2000; Crisp and Burges, 2000; Bennett and Bredensteiner, 2000; Chen et al., 2005; Mavroforakis and Theodoridis, 2006; Barbero et al., 2015). These papers re-parameterize the SVM optimization problem in a way which also provides geometric insights into the SVM solution and makes the tuning parameter more interpretable (roughly controlling the number of support vectors). Section 3.7.2 discusses how we can use the nu-SVM formulation and our results to provide additional insights into SVM. The nu-SVM formulation could also be used to prove some of our results (e.g. a weaker version of parts of Theorem 3.4.5) in a different way (however, we believe our proof techniques require less background work). The nu-SVM literature is mostly focused on computation and we did not find much overlap with our results.

SVM robustness properties have been previously studied ((Schölkopf et al., 2000; Steinwart and Christmann, 2008)), however, the cropped MD characterization of SVM for small $C$ appears to be new.

We find the gap and diameter (Definitions 3.4.2, 3.4.1) of the dataset are important quantities for SVM tuning. These quantities show up in other places in the SVM literature, for example, their ratio is an important quantity in statistical learning theory (Vapnik, 1999).

Some previous papers have suggested modifying SVM's intercept (Crisp and Burges, 2000). We suggest a particular modification (Section 3.6.2) which addresses the *margin bounce* phenomena (Section 3.5.3).

SVM tuning has been extensively studied ((Steinwart and Christmann, 2008)[Chapter 11]). Some papers focus on computational aspects of SVM tuning e.g. cheaply computing the full tuning path (Hastie et al., 2004). Other papers focus on tuning kernel parameters (Sun et al., 2010). Some papers optimize alternative metrics which attempt to better approximate the test set error (Chapelle and Vapnik, 2000; Ayat et al., 2005). Some papers propose default values for tuning parameters (Mattera and Haykin, 1999;

Cherkassky and Ma, 2004). Our tuning results provide different kinds of insights whose applications are discussed in more detail in Section 3.6.

### 3.2 Setup and Notation

A linear classifier is defined via the *normal vector* to its discriminating hyperplane and an *intercept* (or *offset*). A key idea in this chapter is to compare *directions* of linear classifiers. Comparing the direction between two classifiers means comparing their normal vector directions; we say two directions are equivalent if one is a scalar multiple of the other (see Section 2). Note that two classifiers may have the same direction, but lead to different classification algorithms (i.e. the intercepts may differ).

Suppose we have $n$ labeled data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and index sets $I_+, I_-$ such that $y_i = 1$ if $i \in I_+$, $y_i = -1$ if $i \in I_-$ and $\mathbf{x}_i \in \mathbb{R}^d$. Let $n_+ = |I_+|$ and $n_- = |I_-|$ be the class sizes (i.e. $n_- + n_+ = n$). We consider linear classifiers whose decision function is given by

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b,$$

where $\mathbf{w} \in \mathbb{R}^d$ is the normal vector and $b \in \mathbb{R}$ is the intercept (classification rule $\text{sign}(f(x))$).

Given two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ we consider their *directions to be equivalent* if there exists $a \in \mathbb{R}, a \neq 0$ such that $a\mathbf{w} = \mathbf{v}$ (and we will write $\mathbf{w} \propto \mathbf{v}$). Using this equivalence relation we can quotient $\mathbb{R}^d$ into the space of directions (formally real projective space). Intuitively, this is the space of lines through the origin.

In this chapter we consider the following linear classifiers: hard margin SVM, soft margin SVM (which we refer to as SVM), mean difference (also called *nearest centroid*), and the maximal data piling direction.

Often linear classification algorithms can be extended to a wide range of non-linear classification algorithms using the *kernel trick* (Schölkopf and Smola, 2002). While a kernlized linear classifier is no longer linear in the original data, it is a linear classifier in some transformed space (often called the *feature space*). Therefore, in this chapter we focus on the linear case, but our mathematical results extend to the kernel case.

### 3.2.1 Mean Difference and Convex Classifiers

The *mean difference* (MD) classifier selects the hyperplane that lies half way between the two class means. In particular the vector $\mathbf{w}_{md}$ is given by the difference of the class means

$$\begin{aligned} \mathbf{w}_{md} &:= \frac{1}{n_+}\sum_{i\in I_+}\mathbf{x}_i - \frac{1}{n_-}\sum_{i\in I_-}\mathbf{x}_i \\ &:= \bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-. \end{aligned} \tag{3.2.1}$$

By replacing the mean with another measure of center (e.g. the *spatial median* Brown 1983) we can motivate a number of other classifiers.

We say a linear classifier is a *convex classifier* if its normal vector, $\mathbf{w}$ is given as the difference of points lying in the convex hulls of the two classes (i.e. $\mathbf{w} = \mathbf{c}_+ - \mathbf{c}_-$ where $\mathbf{c}_\pm \in \text{conv}(\{\mathbf{x}_i | i \in I_\pm\})$). These classifiers are sometimes refered to as *nearest centroid* classifiers because they classify test points by assigning them to the class with the nearest centroid, $\mathbf{c}_+$ or $\mathbf{c}_-$ .

We define *convex directions*, $C$, to be the set of directions such a classifier can take.

**Definition 3.2.1.** *Let $C$ denote the set of all vectors associated with the directions that go between the convex hulls of the two classes i.e.*

$$C := \{a(\mathbf{c}_+ - \mathbf{c}_-) \,|\, a \in \mathbb{R}, a \neq 0, \text{ and } \mathbf{c}_j \in \text{conv}(\{\mathbf{x}_i\}_{i\in I_j}), j = \pm\}.$$

The set $C$ may be all of $\mathbb{R}^d$ if, for example, the two convex hulls intersect. When the data are linearly separable $C$ is a strict subset of $\mathbb{R}^d$. This set of directions will play an important role in later sections.

### 3.2.2 Data Transformation

It is common to transform the data before fitting a linear classifier, for example, the analyst may mean center the variables then scale them by the standard deviation. A number of classifiers can be viewed as either: apply a data transformation then fit a more simple classifier (such as MD) or as a distinct classifier. These classifiers include: *naive Bayes*, *Fisher linear discrimination*, *nearest shrunken centroid*, *regularized discriminant analysis*, and more (Friedman et al., 2001).

For example, when $d < n-1$ the Fisher's linear discriminant direction is given by

$$\mathbf{w}_{fld} := \widehat{\Sigma}_{pool}^{-1}(\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-), \tag{3.2.2}$$

letting $X_-$ and $X_+$ be the data matrix for the respective classes and the *pooled sample covariance* is $\widehat{\Sigma}_{pool} := \frac{1}{n-2}\left[(X_+ - \overline{X}_+)^T(X_+ - \overline{X}_+) + (X_- - \overline{X}_-)^T(X_- - \overline{X}_-)\right]$. Note the inevitability of $\widehat{\Sigma}_{pool}$ plays an important role in the next section.

It is easy to see FLD is equivalent to transforming the data by the pooled sample covariance matrix (i.e. multiplied each data point by $\widehat{\Sigma}_{pool}^{-1/2}$) then computing the MD classifier (where we apply the same transformation to the test data). More generally, if we have a simple, convex classifier (e.g. the MD) given by $\mathbf{w}$ and we apply a data transformation in the form of $\Sigma^{-1/2}$ to the data we obtain the same classifier as $\Sigma^{-1}\mathbf{w}$.

The technical results of this chapter connect SVM to MD (and various other convex classifiers), however, they apply more generally. If the analyst first transforms the data before fitting SVM, as is common in practice, then our results connect SVM to the more general classifier. For example, naive Bayes is equivalent to first transforming the data by a certain diagonal covariance matrix; in this case, our results connect SVM to naive Bayes.

### 3.2.3 Maximal Data Piling Direction

For linear classifiers one frequently projects the data onto the one dimensional subspace spanned by the normal vector. *Data piling*, first discussed by (Marron et al., 2007), is when multiple points have the same projection on the line spanned by the normal vector. For example, all points on SVM's margin have the same image under the projection map. (Ahn and Marron, 2010b) showed that when $d \geq n-1$ there are directions such that each class is projected to a single point i.e. there is *complete data piling*.

**Definition 3.2.2.** *A vector $\mathbf{w} \in \mathbb{R}^d$ gives complete data piling for two classes of data if there exist $a, b \in \mathbb{R}$, with $a \neq 0$ such that*

$$\mathbf{w}^T\mathbf{x}_i = ay_i + b \text{ for each } i = 1, \ldots, n,$$

*where $b$ is the midpoint of the projected classes and $a$ is half the distance between the projected classes.*

The *maximal data piling* (MDP) direction, as its name suggests, searches around all directions of complete data piling and finds the one that maximizes the distance between the two projected class images. This classifier has been studied in a number of papers such as (Ahn et al., 2012), (Lee et al., 2013), and (Ahn and Marron, 2010b). The MDP direction can be computed analytically

$$\mathbf{w}_{mdp} = \widehat{\Sigma}^{-}(\bar{\mathbf{x}}_{+} - \bar{\mathbf{x}}_{-}), \tag{3.2.3}$$

where $A^{-}$ is the Moore-Penrose inverse of a matrix $A$ and $\widehat{\Sigma} := \frac{1}{n-1}(X - \bar{X})^{T}(X - \bar{X})$ is the *global sample covariance* matrix (in contrast with the pooled sample covariance of FLD given above).

The MDP direction has an interesting relationship to Fisher linear discrimination. Recall the formula for FLD show in Equation 3.2.2 above. (Ahn and Marron, 2010b) showed that in low dimensional settings FLD and the MDP formula are the same (though in low dimensional settings MDP does not give complete data piling); when $d < n-1$ the above two equations are equivalent.

Another view of this relation comes from the optimization perspective. FLD attempts to find the direction that maximizes the ratio of the projected "between-class variance to the within-class variance," (Bishop, 2006). This problem is well defined only in low dimensions; in high dimensions when $d \geq n-1$ there exist directions of complete data piling where the within class projected variance is zero. In the high dimensional setting MDP searches around these directions of zero within class variance to find the one that maximizes the distance between the two classes (i.e. the between-class variance).

### 3.2.4 Support Vector Machine

Hard margin support vector machine is only defined when the data are linearly separable; it seeks to find the direction that maximizes the margin separating the two classes. It is defined as the solution to the following optimization problem,

$$\begin{aligned} \underset{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to} \quad & y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1, \text{ for } i = 1, \ldots, n. \end{aligned} \tag{3.2.4}$$

When the data are not separable Problem (3.2.4) can be modified to give soft margin SVM by adding a tuning parameter $C$ and slack variables $\xi_i$ which allow points to be on the wrong side of the margin,

$$\begin{aligned} \underset{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i \\ \text{subject to} \quad & y_i(\mathbf{x}_i\cdot\mathbf{w}+b) \geq 1-\xi_i, \text{ for } i=1,\ldots,n \\ & \xi_i \geq 0, \text{ for } i=1,\ldots,n. \end{aligned} \tag{3.2.5}$$

For a detailed introduction to SVM see (Mohri et al., 2012).

In both cases the direction is a linear combination of the training data points

$$\mathbf{w}_{svm} = \sum_{i\in I_+} \alpha_i \mathbf{x}_i - \sum_{i\in I_-} \alpha_i \mathbf{x}_i.$$

It turns out this linear combination always gives a direction that points between the convex hull of the two classes (see Definition 3.2.1).

## 3.3 Hard Margin SVM in High Dimensions

In this section we provide novel insights into the geometry of complete data piling which are then used to characterize the relationship between hard margin SVM and MDP in high dimensions. The results are stated in the first two subsections then proved in the remaining two subsection and appendix.

For this section we assume $d \geq n-1$. We further assume the data are in general position and separable, which implies the data are linearly independent if $d \geq n$ and affine independent if $d = n-1$. The data are in general position with probability 1 if they are generated by an absolutely continuous distribution in high dimensions. Typically the phenomena studied here happens in the $n-1$ dimensional affine space generated by the data.

### 3.3.1 Complete Data Piling Geometry

Define the set $P$ of *complete data piling directions* using ideas from Definition 3.2.2.

**Definition 3.3.1.** *Let $P$ denote the vectors associated with directions that give complete data piling i.e.*

$$P := \{\mathbf{v}\in\mathbb{R}^d | \exists a, b\in\mathbb{R}, a\neq 0 \text{ s.t. } \mathbf{v}^T\mathbf{x}_i = a\cdot y_i + b \text{ for each } i=1,\ldots,n\}.$$

Note the set of complete data piling directions can be empty, however, if the data are in general position then $P \neq \emptyset$ when $d \geq n-1$. In this case, (Ahn and Marron, 2010b) point out there are infinitely many of such directions in the ($n$ dimensional) subspace generated by the data that give complete data piling; in fact there is a great circle of directions in this subspace (if we parameterize directions by points on the unit sphere).

Theorem 3.3.2 shows there is a single complete data piling direction that is also within the ($n-1$ dimensional) affine hull of the data. The remaining directions in $P$ come from linear combinations of this unique direction in the affine hull and any vector normal to that hull.

**Theorem 3.3.2.** *The set of complete data piling directions, $P$, intersects the affine hull of the data in a single direction which is the maximal data piling direction.*

Theorem 3.3.2 is proved in the appendix.

### 3.3.2 Hard Margin SVM and Complete Data Piling

A simple corollary of Theorem 3.3.2 is:

**Corollary 3.3.3.** *The intersection of the convex directions, $C$, and the complete data piling directions, $P$, is either empty or a single direction i.e.*

$$C \cap P = \emptyset \text{ or } C \cap P = \{a\mathbf{v} | a \in \mathbb{R}\}.$$

In other words, if a convex classifier gives complete data piling then it has to also be the MDP; furthermore, there can be at most one convex classifier which gives complete data piling.

The core results for hard margin SVM are summarized in the following theorem. Note that this theorem also characterizes when SVM has complete data piling

**Theorem 3.3.4.** *The hard margin SVM and MDP directions are equivalent if and only if there is a non-empty intersection between the convex directions, $C$, and the complete data piling directions, $P$. In this case, the intersection is a single direction which is the hard margin SVM direction and the MDP direction i.e.*

$$\mathbf{w}_{hm-svm} \propto \mathbf{w}_{mdp} \Leftrightarrow P \cap C \neq \emptyset \Leftrightarrow \mathbf{w}_{hm-svm} \propto \mathbf{w}_{mdp} = C \cap P$$

Where we use the equality sign to indicated $C \cap P$ is a single direction. Theorem 3.3.4 is a consequence of Corollary 3.3.3, Lemma 3.3.7, Lemma 3.3.8 and the KKT conditions.

Appendix 3.7.3 gives an alternate characterization of the event $P \cap C \neq \emptyset$ through a linear program which is of theoretical interest.

As a corollary of this theorem we can characterized when MD/MDP or SVM/MD are equivalent.

**Corollary 3.3.5.** *The hard margin SVM and MD directions are equivalent if and only if all three of hard margin SVM, MD and MDP are equivalent i.e.*

$$\mathbf{w}_{hm-svm} \propto \mathbf{w}_{md} \Leftrightarrow \mathbf{w}_{md} \propto \mathbf{w}_{mdp}.$$

Another corollary of this theorem is that hard margin SVM is always the MDP of the support vectors.

**Corollary 3.3.6.** *Let $V$ be the set of support vectors for hard margin SVM, then $\mathbf{w}_{hm-svm}$ is the MDP of $V$.*

This Corollary says that we can interpret hard margin SVM as a cropped MDP (i.e. it ignores points which are far away from the separating hyperplane).

### 3.3.3 Hard Margin KKT Conditions

Derivation and discussion of the KKT conditions can be found in (Mohri et al., 2012). From the Lagrangian of Problem (3.2.4) we can derive the KKT conditions

$$\mathbf{w}_{hm-svm} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \tag{3.3.1}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \tag{3.3.2}$$

$$\alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1, \tag{3.3.3}$$

with $\alpha_i \geq 0$ for each $i = 1, \ldots, n$.

Condition (3.3.2) says that the sum of the weights in both classes has to be equal. Combining this with (3.3.1) we find that the hard margin SVM direction is given by

$$\mathbf{w}_{hm-svm} \propto \sum_{i \in I_+} \frac{\alpha_i}{A} \mathbf{x}_i - \sum_{i \in I_-} \frac{\alpha_i}{A} \mathbf{x}_i, \tag{3.3.4}$$

where $\sum_{i\in I_+}\alpha_i = \sum_{i\in I_-}\alpha_i := A$. Thus $\mathbf{w}_{hm-svm} \in C$ i.e. the hard margin SVM direction is always a convex direction. As discussed in (Bennett and Bredensteiner, 2000; Pham, 2010) hard margin SVM is equivalent to finding the nearest points in the convex hulls of the two classes.

The last KKT condition (3.3.3) says that a point $\mathbf{x}_i$ either lies on one of the marginal hyperplanes $\{\mathbf{x}|\mathbf{w}_{hm-svm}^T\mathbf{x} = \pm 1\}$ or receives zero weight. In the former case when $\alpha_i \neq 0$, $\mathbf{x}_i$ is called a *support vector*.

The margin $\rho$ is defined as the minimum distance from a training point to the separating hyperplane; $\rho$ is also the orthogonal distance from the marginal hyperplanes to the separating hyperplane. The margin width is given by the magnitude of the normal vector

$$\rho^2 = \frac{1}{||\mathbf{w}_{hm-svm}||_2^2} = \frac{1}{\sum_{i=1}^n \alpha_i} := \frac{1}{||\alpha||_1}. \tag{3.3.5}$$

### 3.3.4 Proofs for Hard Margin SVM

The following lemma about SVM and MDP is a consequence of the fact that complete data piling directions satisfy the SVM KKT conditions.

**Lemma 3.3.7.** *If hard margin SVM has complete data piling then the SVM direction is equivalent to the MDP direction i.e.*

$$\mathbf{w}_{hm-svm} \in P \implies \mathbf{w}_{hm-svm} \propto \mathbf{w}_{mdp}.$$

**Lemma 3.3.8.** *If* $P \cap C \neq \emptyset$ *then* $\mathbf{w}_{svm} \in P \cap C$.

*Proof.* Let $\mathbf{v} \in P \cap C$. We show $\mathbf{v}$ satisfies the KKT conditions. The lemma then follows since the KKT conditions necessary and sufficient for hard margin SVM (the constraints are qualified, see Chapter 4 of Mohri et al. 2012).

Since $\mathbf{v} \in C$ we have that $\mathbf{v} \propto \mathbf{c}_+ - \mathbf{c}_-$ where $\mathbf{c}_j \in \text{conv}(\{\mathbf{x}_i\}_{i\in I_j})$. For some constant $a > 0$

$$\mathbf{v} = a\left(\sum_{i\in I_+}\lambda_i\mathbf{x}_i - \sum_{i\in I_-}\lambda_i\mathbf{x}_i\right),$$

where

$$\sum_{i\in I_+}\lambda_i = \sum_{i\in I_+}\lambda_i = 1 \text{ and } \lambda_i \geq 0.$$

Since $\mathbf{v} \in P$ we can select $b, \mathbf{v}$ such that

$$y_i(\mathbf{x}_i \cdot \mathbf{v} + b) = 1 \ \forall i.$$

But these three equations are the KKT conditions with $\alpha_i = a\lambda_i$.

■

### 3.4 Soft Margin SVM Small and Large $C$ Regimes

This section characterizes the behavior of SVM for the small and large regimes of the cost parameter $C$. We make no assumptions about the dimension of the data $d$. We state the main results for the small and large $C$ regimes, provide the KKT conditions, then prove the tuning regimes results.

We first make two geometric definitions that play an important role in characterizing SVM's tuning behavior. The two class *diameter* measures the spread of the data.

**Definition 3.4.1.** *Let the two class diameter be*

$$D := \max_{\mathbf{x}_+ \in I_+, \mathbf{x}_- \in I_-} ||\mathbf{x}_+ - \mathbf{x}_-||.$$

The *gap* measures the separation between the two data classes.

**Definition 3.4.2.** *Let the two class gap $G$ be the minimum distance between points in the convex hulls of the two classes i.e.*

$$G := \min_{\mathbf{c}_j \in conv(\{\mathbf{x}_i\}_{i \in I_j})} ||\mathbf{c}_+ - \mathbf{c}_-||.$$

If the data are not linearly separable then $G = 0$.

Using the above geometric quantities we define two threshold values of $C$ which determine when the SVM enters its different behavior regimes.

**Definition 3.4.3.** *For two classes of data let*

$$C_{small} := \frac{2}{\max(n_+, n_-)D^2}, \tag{3.4.1}$$

*where $D$ is the diameter of the training data.*

**Definition 3.4.4.** *If the two data classes are linearly separable let*

$$C_{large} := \frac{2}{G^2}, \tag{3.4.2}$$

*where $G$ is the gap between the classes.*

As illustrated in Figures 3.1.2 and 3.1.3, the main result for the small $C$ regime is given by Theorem 3.4.5 and Corollary 3.4.7. We call the support vectors lying strictly within the margin *slack vectors* (Definition 3.4.10).

**Theorem 3.4.5.** *When every point in the smaller (negative) class is a slack vector,*

- *if the classes are balanced then the SVM direction becomes the mean difference direction i.e. $\mathbf{w}_{svm} \propto \mathbf{w}_{md}$.*
- *if the classes are unbalanced then the SVM direction satisfies the constraints in Equations 3.4.3, 3.4.4 making it a cropped mean difference.*

$$\mathbf{w}_{svm} = \sum_{i \in M_+} \alpha_i \mathbf{x}_i + C \sum_{i \in L_+} \mathbf{x}_i - C \sum_{i \in I_-} \mathbf{x}_i, \tag{3.4.3}$$

*subject to*

$$\sum_{i \in M_+} \alpha_i = C(|L_+| - n_-). \tag{3.4.4}$$

*Furthermore, $C < C_{small}$ is a sufficient condition such that every point in the smaller class is a slack vector.*

Theorem 3.4.5 characterizes a kind of cropped mean difference. The mean difference direction points between the mean of the first class and the mean of the second class. Recall $\mathbf{w}_{svm}$ always goes between points in the convex hulls of the two classes. Equation 3.4.3 says that in the small $C$ regime $\mathbf{w}_{svm}$ points between the mean of the smaller (negative) class (the third term) and a point that is close to the mean in the larger (positive) class. The cropping happens by ignoring non-support vectors. While points on the margin do not necessarily receive equal weight, Equation 3.4.4 bounds the amount of weight put on points on margin points. Note Equations 3.4.3, 3.4.4 are stronger than the simple constraint that $\sum_{i \in I_+} \alpha_i = n_- C$ (Lemma 2 from (Hastie et al., 2004)) since all of the slack vectors in the positive class receive the same weight.

Lemma 3.4.6 strengthens Lemma 3.4.5 in the case $n_+ >> d$ (i.e. there can't be too many margin vectors in Equation 3.4.3 )

**Lemma 3.4.6.** *If the data are in general position the larger class can have at most $n_- + d - 1$ support vectors.*

As $C$ continues to shrink past $C_{\text{small}}$ the margin width continues to grow. Eventually the separating hyperplane will be pushed past the smaller class and every training point will be classified to the larger class (see Figure 3.1.3d). Note this results follows from the proofs in Section 3.4.2.

**Corollary 3.4.7.** *If the classes are unbalanced and $C < \frac{1}{2} C_{small}$ then every training point is classified to the larger (positive) class.*

If the data are separable then in the large $C$ regime soft margin SVM becomes equivalent to hard margin SVM for sufficiently large $C$.

**Theorem 3.4.8.** *If the training data are separable then when $C > C_{large}$, soft margin SVM is equivalent to the hard margin SVM solution i.e. $\mathbf{w}_{svm} = \mathbf{w}_{hm-svm}$.*

Note that $C_{\text{small}}$ and $C_{\text{large}}$ are lower and upper bounds—their respective limiting behavior may happen for $C$ larger that $C_{\text{small}}$ and $C$ smaller than $C_{\text{large}}$. In practice, these threshold values are a reasonable approximation. Furthermore, the $\frac{1}{D^2}$ scales is important for small values of $C$ (this can be seen in the proofs of Corollary 3.4.13 and Lemma 3.4.14)

### 3.4.1 Soft Margin SVM KKT Conditions

The KKT conditions[1] for soft margin SVM are (see Mohri et al. 2012 for derivations)

$$\mathbf{w}_{svm} = \sum_{i \in I_+} \alpha_i \mathbf{x}_i - \sum_{i \in I_-} \alpha_i \mathbf{x}_i, \tag{3.4.5}$$

$$\sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i := A, \tag{3.4.6}$$

$$\alpha_i + \mu_i = C \text{ for } i = 1, \ldots, n, \tag{3.4.7}$$

[1] Note "or" should be read as non-exclusive or i.e. both may be true.

$$\alpha_i = 0 \text{ or } y_i(\mathbf{w}\cdot\mathbf{x}_i + b) = 1 - \xi_i \text{ for } i = 1,\ldots,n, \tag{3.4.8}$$

$$\xi_i = 0 \text{ or } \mu_i = 0 \text{ for each } i, \tag{3.4.9}$$

For soft margin SVM we define the marginal hyper planes to be $\{\mathbf{x}|\mathbf{x}^T\mathbf{w}_{svm} = \pm 1\}$ and the margin width (or just margin), $\rho$ the distance from the separating hyperplane to the marginal hyperplanes. By construction $\rho = \frac{1}{||\mathbf{w}_{svm}||}$. For soft margin SVM, the margin does not have the same meaning as in the hard margin case, but still plays an important role. In particular, a points is a support vector if and only if it is contained within the marginal hyperplanes.

As with hard margin SVM, the soft margin direction is always a convex direction. Again points $\mathbf{x}_i$ such that $\alpha_i \neq 0$ are called support vectors. We further separate support vectors into two types.

**Definition 3.4.9.** *Margin vectors are support vectors* $\mathbf{x}_i$ *such* $\alpha_i \neq 0$ *and* $\xi_i = 0$.

**Definition 3.4.10.** *Slack vectors are support vectors* $\mathbf{x}_i$ *such* $\alpha_i \neq 0$ *and* $\xi_i > 0$.

Margin vectors are support vectors lying on one of the two marginal hyperplanes. Slack vectors are support vectors lying strictly on the inside of the marginal hyperplanes. Call the set of margin vectors in each class $M_j$ and the set of slack vectors $L_j$ for $j = \pm$.

The KKT conditions imply

- all support vectors receive weight upper bounded by $C$ ($\mathbf{x}_i \in M_j \implies 0 < \alpha_i \leq C$)
- slack vectors receive weight exactly $C$ ($\mathbf{x}_i \in L_j \implies \alpha_i = C$)

Furthermore, the following constraint balances the weights between the two classes

$$C|L_+| + \sum_{i\in M_+} \alpha_i = C|L_-| + \sum_{i\in M_-} \alpha_i. \tag{3.4.10}$$

We assume that the positive class is the larger of the two classes i.e. $n_+ \geq n_-$. Unbalanced classes means $n_+ > n_-$.

### 3.4.2 Proofs for Small $C$ Regime

As $C \to 0$ the margin width increases to infinity ($\rho \to \infty$). As the margin width grows as many points as possible become slack vectors and all slack vectors get the same weight $\alpha_i = C$. Hence if the classes are balanced the SVM direction will be equivalent to the mean difference. If the classes are unbalanced then there will be some margin vectors which receive weight $\alpha_i \leq C$. The number of margin vectors is bounded by the class sizes and the dimension.

Note the diameter, $D$, does not change if we consider the convex hull of the two classes (proof of Lemma 3.4.11 is a straightforward exercise).

**Lemma 3.4.11.**

$$\max_{\mathbf{c}_j \in conv(\{\mathbf{x}_i\}_{i \in I_j})} ||\mathbf{c}_+ - \mathbf{c}_-|| = \max_{\mathbf{x}_j \in I_+} ||\mathbf{x}_+ - \mathbf{x}_-|| =: D.$$

As $C \to 0$ the magnitude of $\mathbf{w}_{svm}$ goes to zero. In particular, the KKT conditions give the following bound.

**Lemma 3.4.12.** *For a given $C$ the magnitude of the SVM solution is*

$$||\mathbf{w}_{svm}|| \leq n_+ C \cdot D.$$

*Proof.* From the KKT conditions we have

$$\mathbf{w}_{svm} = \sum_{i \in I_+} \alpha_i \mathbf{x}_i - \sum_{i \in I_-} \alpha_i \mathbf{x}_i$$

and

$$\sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i =: A.$$

Computing the magnitude of $\mathbf{w}_{svm}$

$$||\mathbf{w}_{svm}|| = A \left|\left| \sum_{i \in I_+} \frac{\alpha_i}{A} \mathbf{x}_i - \sum_{i \in I_-} \frac{\alpha_i}{A} \mathbf{x}_i \right|\right|.$$

Since the two terms are convex combinations we get

$$||\mathbf{w}_{svm}|| \leq A \sup_{\mathbf{c}_j \in \text{conv}(\{\mathbf{x}_i\}_{i \in I_j})} ||\mathbf{c}_+ - \mathbf{c}_-||.$$

applying Lemma 3.4.11

$$||\mathbf{w}_{svm}|| = A \max_{\mathbf{x}_j \in I_+} ||\mathbf{x}_+ - \mathbf{x}_-||$$

$$||\mathbf{w}_{svm}|| = AD.$$

Since $0 \leq \alpha_i \leq C$ we get $A \leq n_1 C$ thus proving the bound. ■

Since the magnitude of $\mathbf{w}_{svm}$ determines the margin width, using the previous lemma we get the following corollary.

**Corollary 3.4.13.** *The margin $\rho$ goes to infinity as $C$ goes to zero. In particular*

$$\rho = \frac{1}{||\mathbf{w}_{svm}||} \geq \frac{1}{n_+ CD}.$$

Since the margin width increases, for small enough $C$ the smaller class becomes all slack variables.

**Lemma 3.4.14.** *If $C < C_{small}$ then all points in the smaller class become slack vectors ($\xi_i > 0$ for all $i \in I_-$).*

*Proof.* By Corollary 3.4.13 the margin width goes to infinity as $C \to 0$ since

$$\rho \geq \frac{1}{n_+ CD}.$$

Recall the margin width, $\rho$, is the distance from the separating hyperplane to the marginal hyperplanes. Note that if $\rho > \frac{1}{2}D$ then at least one class must be complete slack. Thus if $C < \frac{2}{n_1 D^2}$ at least one class must be complete slack i.e. $\xi_i > 0$ for all $i \in I_j$ for $j = +$ and/or $j = -$. If the classes are balanced then either class can become complete slack (or both classes).

If the classes are unbalanced i.e. $n_- < n_+$ then the smaller class becomes complete slack. To see this, assume for the sake of contradiction that the larger class becomes complete slack i.e. $\xi_i \neq 0$ for each $i \in I_+$. Then the KKT conditions imply $\alpha_i = C$ for each $i \in I_+$. KKT condition 3.4.6 says

$$\sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i$$

$$n_+ C = \sum_{i \in I_-} \alpha_i.$$

But $\alpha_i \leq C$ and $n_- < n_+$ by assumption therefore this constraint cannot be satisfied. ■

If the classes are balanced then the margin swallows both classes and the SVM direction becomes the mean difference direction.

**Lemma 3.4.15.** *If the classes are balanced and $C < C_{small}$ the SVM direction is equivalent to the mean difference direction i.e. $\mathbf{w}_{svm} \propto \mathbf{w}_{md}$.*

*Proof.* When $C < C_{\text{small}}$ one of the classes (without loss of generality the negative class) becomes slack i.e. $\xi_i > 0$ for each $i \in I_-$ thus $\alpha_i = C$ for each $i \in I_-$. The KKT conditions then require

$$\sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i = n_- C.$$

Since $\alpha_i \leq C$ and $|I_+| = n_-$ this constraint can only be satisfied if $\alpha_i = C$ for each $i \in I_+$. We now have

$$\mathbf{w}_{svm} = \sum_{i \in I_+} C\mathbf{x}_i - \sum_{i \in I_-} C\mathbf{x}_i$$

$$\mathbf{w}_{svm} = C\frac{n}{2}(\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-) \propto \mathbf{w}_{md}.$$

■

**Lemma 3.4.16.** *If the classes are unbalanced and $C < C_{small}$ the SVM solution satisfies the the constraints in Equations 3.4.3, 3.4.4.*

*Proof.* Recall for $C < C_{\text{small}}$ we have $\xi_i > 0$ for $i \in I_-$. From the KKT conditions $\xi_i > 0 \implies \mu_i = 0 \implies \alpha_i = 0$ meaning $\alpha_i = C$ for each $i \in I_-$. The weight balance constraint 3.4.10 from the KKT conditions becomes

$$C|L_+| + \sum_{i \in M_+} \alpha_i = C|L_-| + \sum_{i \in M_-} \alpha_i,$$

which then implies the conditions on $\mathbf{w}_{svm}$. ■

**Corollary 3.4.17.** *When $C < C_{small}$ the larger (positive) class can have at most $n_-$ slack vectors. If the larger class has more than $n_-$ support vectors then at least one of them must be a margin vector.*

### 3.4.3 Proofs for Large $C$ Regime

**Lemma 3.4.18.** *If there is at least one slack vector then for a given $C$*

$$||\mathbf{w}_{svm}|| \geq CG,$$

*or equivalently*

$$\rho \leq \frac{1}{CG},$$

*where $G$ is the class gap.*

*Proof.* From the KKT conditions

$$||\mathbf{w}_{svm}|| = ||\sum_{i \in I_+} \alpha_i \mathbf{x}_i - \sum_{i \in I_-} \alpha_i \mathbf{x}_i||,$$

$$||\mathbf{w}_{svm}|| = A||\sum_{i \in I_+} \frac{\alpha_i}{A} \mathbf{x}_i - \sum_{i \in I_-} \frac{\alpha_i}{A} \mathbf{x}_i||,$$

where $A = \sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i$. Since the two sums are convex combinations, using the definition of $G$ we get

$$||\mathbf{w}_{svm}|| \geq AG.$$

Since there is at least one slack vector there is at least one $i$ such that $\alpha_i = C$ thus $A \geq C$ and the result follows. ■

## 3.5 Summary of SVM Regimes

For sufficiently small values of $C$, SVM is related to the mean difference. When the data are separable, for sufficiently large values of $C$ soft margin SVM is equivalent to hard margin SVM. We note this discussion applies more broadly than just binary, linear SVM. For example, when a kernel is used, SVM becomes related to the kernel mean difference classifier. Often multi-class classification problems are reduced to a number of binary class problems e.g. using *one vs. one* (OVO) or *one vs. all* (OVA) schemes. Our results apply to each of these binary classification problems. For example, in a multi-class problem, even if the classes are roughly balanced, the OVA scheme may produce unbalanced classes where the behavior discussed in Section 3.5.2 becomes applicable.

### 3.5.1 Small $C$ Regime and the Mean Difference

For sufficiently small $C$ (when every point in the smaller class is a slack vector) Theorem 3.4.5 shows how soft margin SVM is related to the mean difference.

If the data are unbalanced then the SVM direction becomes a cropped mean difference direction as characterized by Equations 3.4.3, 3.4.4. The direction points from the mean of the smaller class to a cropped mean of a subset of points in the larger class. The cropped mean of the larger class gives equal weight to slack vectors, puts smaller weight on margin vectors and ignores points that are outside the margin (non-support vectors). Furthermore, the number of margin vectors is bounded by the dimension when the data are in general position (Lemma 3.4.6).

In the small $C$ regime, if the data are balanced then the SVM direction becomes exactly the mean difference direction. Note Lemma 1 from (Hastie et al., 2004) proves this result for balanced classes, proves a weaker version in the unbalanced case, does not give the threshold $C_{\text{small}}$, and does not discuss the connection between SVM and the MD classifier.

The lower bound $C_{\text{small}}$ is important because it shows SVM's MD like behavior applies for every dataset set. Furthermore, it shows that the value of $C$ where the MD like behavior begins depends on the data diameter and class sizes $\left(\text{i.e. is proportional to } \frac{1}{\max(n_+,n_-)D^2}\right)$. This dependence on the data diameter has important consequences for cross-validation which are discussed in Section 3.6.

Note the cropped MD interpretation is often valid for a wide range of $C$ (i.e. values of $C$ larger than $C_{\text{small}}$). In particular, as $C$ shrinks, more vectors become slack vectors receiving equal weight (see proofs and results in Section 3.4). As $C$ shrinks to $C_{\text{small}}$, the angle between SVM and the cropped MD defined in Theorem 3.4.5 approaches zero. This can be seen, for example, in Figure 3.1.3f.

Finally, note that the relation between SVM and the MD also relates SVM to a larger set of classifiers by taking data transformation into account (see Section 3.2.2). It is common to apply a transformation to the data before fitting SVM (e.g. mean centering then scaling by some covariance matrix estimate). In this case, the small $C$ regime of SVM will be a (cropped) version of the transformed MD classifier. This insight connects SVM to, for example, the naive Bayes classifier. Similarly, our results also connect kernel SVM to the kernel (cropped) MD classifier.

SVM's MD behavior discussed in this section raises the question of how much performance gain SVM achieves over (robust, transformed) mean difference classifiers. This is discussed more in Section 3.7.3.

### 3.5.2 Class Imbalance and the MD Regime

Theorem 3.4.5 gives some insights into SVM when the classes are imbalanced. When SVM is in the MD regime as discussed above (i.e. $C \leq C_{\text{small}}$), every point in the smaller (negative) class has to be a support vector receiving equal weight. In some scenarios the MD or a cropped MD may perform very well. However, this result says in the small $C$ regime, SVM cannot crop the smaller class (it can still crop the smaller class when $C > C_{\text{small}}$). This insight can explain some scenarios where SVM performs well for small values of $C$, but then its performance suddenly degrades for even smaller values of $C$ (i.e. an outlier is forced into the smaller class's slack vectors).

Lemma 3.4.6 says that (under weak conditions) the larger (positive) class can have at most $n_- + d + 1$ support vectors ($n_-$ = size of the smaller class). In the case $n_+ >> n_-, d$ then SVM can only use a small number of data points from the larger class to estimate the SVM direction (this is true for all values of C). This means SVM is forced to do a lot of cropping for the larger (positive) class which may be a good thing in some scenarios (i.e. if the larger class has many outliers).

### 3.5.3 Small $C$ Regime and Margin Bounce

As $C$ shrinks, the margin (distance between the marginal hyperplanes) increases. When the classes are unbalanced, the marginal hyperplane of the larger class has to stay within the convex hull of the larger class causing the separating hyperplane to move off to infinity. For small enough values of $C$ ($\leq \frac{1}{2} C_{\text{small}}$), this means the separating hyperplane is pushed past the smaller class and every point is classified to the larger class (Corollary 3.4.7). We call this behavior *margin bounce* (see Figure 3.1.3a for an example). In other words, for small values of $C$, SVM picks a reasonable direction, but a bad intercept.

When the classes are exactly balanced, the margin bounce may or may not happen (we have seen data examples of both). It would be an interesting follow up question to determine conditions for when the margin bounce happens for balanced classes.

This insight has a few consequences.

1. For Figure 3.1.3d (unbalanced classes) it explains why the three tuning error curves are large for small values of $C$.

2. For Figure 3.1.2d (balanced classes) it explains why only the cross-validation error curve is bad for small values of $C$, but the tuning and test set error curves are fine (i.e. the cross-validation training sets are typically unbalanced).

3. For small values of $C$ SVM picks a bad intercept, but a fine direction. We exploit this fact in Section 3.6.2 to develop an improved intercept for SVM

4. The value of $C$ when the margin starts exploding depends on the diameter of the two classes. This has important implications for cross-validation which are discussed in Section 3.6.1

### 3.5.4 Large $C$ Regime and the Hard-Margin SVM

If the data are separable, Theorem 3.4.8 says that for sufficiently large values of $C$, soft margin SVM will be equivalent to hard margin SVM. Note that in high-dimensions (i.e. $d > n$) the data are always separable (as long as they are in *general position*). If the original dataset is non-separable, but a kernel is used the transformed dataset may in fact be separable (for example, if the implicit kernel dimension is larger than $n$).

Furthermore, the value of $C$ above which soft-margin SVM becomes equivalent to hard margin SVM depends on the gap between the two classes (see Definition 3.4.2). This can have important consequences for cross-validation as discussed in Section 3.6.1.

### 3.5.5 Hard-Margin SVM and the (cropped) Maximal Data Piling Direction

In high dimensions, (i.e. $d \geq n-1$) Theorem 3.3.4 gives geometric conditions for when hard margin SVM gives complete data piling i.e. when the SVM direction is equivalent to the MDP direction. Hard margin SVM always has some data piling; support vectors in the same class project to the same point. In this case SVM is the MDP direction of the support vectors. In this sense, hard margin SVM can be viewed as a cropped MDP direction where points away from the margin are ignored.

Complete data piling is a strict constraint and the SVM normal vector can usually wiggle away from the MDP direction to find a larger margin. This raises the question: is complete data piling with hard

margin SVM a probability zero event when the data are generated by an absolutely continuous distribution? We suspect the answer is no: it occurs with positive, but typically small probability. For example consider three points in $\mathbb{R}^2$.

Often data piling may not be desirable e.g. the normal vector may be sensitive to small scale noise artifacts (Marron et al., 2007). Additionally, the projected data have a degenerate distribution since multiple data points lie on top of each other. However there are cases, such as an autocorrelated noise distribution, when the maximal data piling direction performs well, (Miao, 2015).

Corollary 3.3.6 (SVM is the MDP of the support vectors) also gives an alternative characterization of hard margin SVM. Hard margin SVM searches over every subset of the data points which have a nonempty set of complete data piling directions, computes the MDP of each such subset, and selects the direction giving the largest separation. This characterization is mathematically interesting because it says we can *a priori* restrict the hard margin SVM optimization problem, Equation 3.2.4, to search over a finite set of directions (i.e. the complete data piling directions of the subsets of the data). Furthermore, in some cases, the MDP (Equation 3.2.3) can be cheaply computed or approximated. For example, the analyst may use a low rank approximation to $\widehat{\Sigma}^-$ and/or select a judicious subset of data points. In these scenarios, it may make sense to approximate hard margin SVM with the MDP.

### 3.6 Applications of SVM Regimes

There are are number of ways of tuning soft margin SVM including: heuristic choice, random search, Nelder-Mead and cross-validation (Nelder and Mead 1965; Mattera and Haykin 1999; Chapelle and Vapnik 2000; Hsu et al. 2003; Christmann et al. 2005; Steinwart and Christmann 2008). In practice one of the most popular methods is to select $C$ which optimizes the $K$-fold cross-validation error (Friedman et al. 2001; Hsu et al. 2003). Note for very unbalanced classes, the cross-validation error metric can be replaced with other test set error metrics such as F-score, Kappa, precision/recall, balanced error, AUC (Tan et al., 2005). This section focuses on test set error, but the discussion is relevant to these other error metrics. The discussion also focuses on cross-validation, but similar conclusions can be drawn when a fixed validation set is used. Furthermore, these insights also apply to using cross-validation to estimate the true test set error.

### 3.6.1 Tuning SVM via Cross-Validation

Tuning SVM using cross-validation means attempting to estimate the tuning curve of the test set (the green line marked with triangles in Figures 3.1.2d, 3.1.3d) using the tuning curve from cross-validation (the red line marked with circles). It is known that the optimal hyper-parameter settings for the full training set (of size $n$) may differ from the optimal settings for the cross-validation sets (of size $\left(1-\frac{1}{k}\right)n$); for example, the smaller dataset often favors larger values of $C$ (more regularization) (Steinwart and Christmann, 2008).

The results of this chapter give a number of insights into how features of the data cause the cross-validation tuning curve to differ from the test set tuning curve. In particular, we have show that the tuning curve is sensitive to

1. balanced vs. unbalanced classes,
2. the two class diameter $D$,
3. whether or not the classes are separable,
4. whether or not $d \geq n-1$,
5. the gap between the two classes $G$.

Each of these characteristics can change between the full training set and the cross-validation training sets. When the characteristics change, so can SVM's behavior for small and large values of $C$. Therefore SVM may behave differently for the cross-validation folds than for the full training data.

One dramatic example of this change in behavior can be seen in Figure 3.1.2d as discussed in Sections 3.5.3, and 3.1.1. In this case, the full dataset is balanced, but the cross-validation folds are typically unbalanced.

Another example of tuning behavior differences between the training and cross-validation data can be seen by looking carefully at Figure 3.1.3d. In this figure we can see the cross-validation error rate shoots up for larger values of $C$ than the train/test error rates. The error increases dramatically for small values of $C$ because of the margin bounce phenomena discussed in Section 3.5.3. The value of $C_{\text{small}}$ that guarantees this behavior is a function of the two class diameter $D$ (see Definition 3.4.3). Since there are

fewer points in the cross-validation training set, the diameter is smaller meaning the value of $C_{\text{small}}$ is larger causing the margin to explode for larger values of $C$.

Different data domains in terms of $n << d, n \sim d$, and $n >> d$ can make the above characteristics more or less sensitive to change induced by subsampling. For example, if $n >> d$ then subsampling is least likely to change whether $d \geq n-1$ or significantly modify the diameter $D$. With a kernel, however, even if the original $n >> d$ then it may no longer be true that $n >> d_{\text{implicit}}$ where $d_{\text{implicit}}$ is the dimension of the implicit kernel space. An interesting, possible exception to this was given by (Rahimi and Recht, 2008) where $d_{\text{implicit}}$ may be small.

When $n$ is larger than $d$, but not by much, then subsampling is likely to change whether or not $d \geq n-1$ and whether or not the data are separable. In this case the full training data may not be separable, but the cross-validation sets may be. This means large values of $C$ will cause soft margin SVM to become hard margin SVM for cross-validation, but never for the full training data. This could result in the SVM direction being very different between cross-validation and training.

When $d \geq n-1$ soft margin SVM will become hard margin SVM for $C \geq C_{\text{large}}$ which depends on the gap $G$ between the two classes. Subsampling the data will cause this gap to increase meaning $C_{\text{large}}$ decreases. In this case the hard margin behavior will occur for smaller values of $C$ in the cross-validation sets than for the full training set.

It is desirable to perform cross-validation in a way that is least likely to change some of the above characteristics between the full and the cross-validation training data set. For example,

- If the full training data are balanced one should ensure the cross-validation training classes are also balanced.
- Cross-validation with a large number of folds (e.g. leave one out CV) is least likely to modify the above characteristics of the data.
- When $n > d$ it could be judicious to make sure that $n_{cv} > d$ for each cross-validation training set.
- (Chapelle and Vapnik, 2000) (Section 4) suggests re-scaling the data using the covariance matrix. The analyst may modify this idea by additionally rescaling each cross-validation training set such that the diameter is (approximately) the same as the diameter of the full training set.

- Previous papers have proposed default values for $C$ based on the given dataset (Mattera and Haykin, 1999; Cherkassky and Ma, 2004). Our results suggest other default values in the interval $[C_{\text{small}}, C_{\text{large}}]$ (when the latter exists) may be reasonable. Furthermore, default values which lie in the middle of this range may be preferable. For example, the analyst may try a simple MD classifier (producing similar results to a small $C$), one moderate and one large value of $C$ for SVM.

### 3.6.2 Improved SVM Intercept for Cross-Validation

As discussed in Section 3.5.3, SVM's intercept can be problematic for small values of $C$; for small values of $C$ the margin bounce causes every point to be classified to the larger of the two classes. This fact alone may not be concerning, however, as Theorem 3.4.5 and Definition 3.4.3 show, SVM can behave differently, as a function of $C$, for cross-validation and on the full data set. The subsampled data sets for cross-validation will have a smaller diameter, $D$, meaning the threshold $C_{small}$ is larger for these datasets than for the full dataset. In particular, the margin explosion happens a larger value of $C$ during cross-validation than it does for the full dataset. This will cause the cross-validation test set error to be large for values of $C$ where the test set error may in fact be small.

We can fix this issue by modifying the SVM intercept as follows. Note that previous papers have suggesting modifying SVM's intercept (Crisp and Burges, 2000). Suppose we fit SVM to a dataset and it returns normal vector and intercept $\mathbf{w}_{svm}$ and $b_{svm}$ respectively. Furthermore, define the *SVM centroids* by

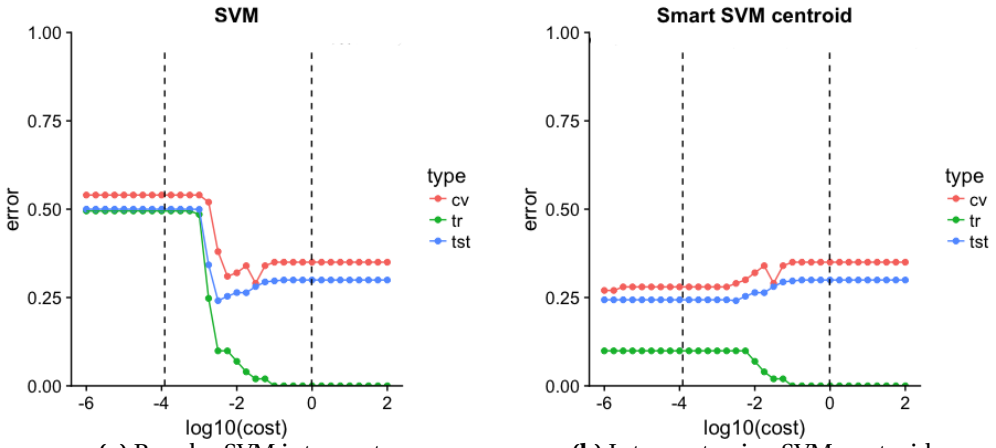
$$\mathbf{m}_{svm,+} = \frac{1}{A}\sum_{i\in I_+} \alpha_i \mathbf{x_i},$$

where the $\alpha_i$ are the support vectors weights and $A$ is the total weight (Equation 3.4.6). Note this is a convex combination of points in the positive class (hence the name SVM centroid). We define $\mathbf{m}_{svm,-}$ similarly for the negative class.

Next define an new intercept by

$$b_{centroid} := \frac{1}{2}\mathbf{w}_{svm}^T(\mathbf{m}_{svm,+} + \mathbf{m}_{svm,-}) \tag{3.6.1}$$

Note $b_{centroid}$ is the value such that SVM's separating hyperplane sits halfway between $\mathbf{m}_{svm,+}$ and $\mathbf{m}_{svm,-}$. Furthermore, note this quantity can be computed when a kernel is used.




**(a)** Regular SVM intercept **(b)** Intercept using SVM centroids

**Figure 3.6.1:** Tuning error curves for standard SVM intercept vs. improved SVM intercept.

The SVM intercept is only a problem when $C$ is small and one class is entirely support vectors (i.e. $\alpha_i > 0 \forall\, i \in I_+$ or $\forall\, i \in I_-$). Finally, we define a new intercept as follows

$$b = \begin{cases} b_{centroid}, & \text{if one class is entirely support vectors} \\ b_{svm}, & \text{otherwise} \end{cases} \tag{3.6.2}$$

Note that when the optimal value of $C$ is large, the margin explosion discussed in this section is not an issue and $b$ defined above will give the same result as the original $b_{svm}$.

The intercepts $b_{centroid}$ and $b$ defined above are not the only options. One could, for example, replace the SVM centroids with the class means (i.e. replace $\mathbf{m}_{svm,-}$ with $\mathbf{x}_+$ ). Alternatively, one could use cross-validation to select $b$ separately from $\mathbf{w}$. We focus on $b_{centroid}$ because it is simple can be interpreted as viewing SVM as a nearest centroid (as discussed in Section 3.2.1).

Below we demonstrate an example where $b$ defined above improves SVM's test set performance. In this example, there are $n_+ = 51$ and $n_- = 50$ points in each class living in $d = 100$ dimensions. The two classes are generated from Gaussians with identity covariance and means which differ only in the first coordinate; the mean of the positive class is the first standard basis vector and the mean of the negative class is negative the first standard basis vector. Note that MD is the Bayes rule in this example. We tune SVM using using 5-fold cross-validation to select the optimal value of $C$ the compute the resulting test set error for an independent test set of 2000 points.

Figure 3.6.1 shows the error tuning curves (as in Figure 3.1.2d) for the two choices of SVM intercepts for a single draw of the data. The x-axis is the tuning parameter and the y-axis is the resulting SVM error for training, testing, and 5-fold cross-validation test set error. In the left panel we see each error curve jumps up to around 50% for small values of $C$ for the regular SVM intercept. Furthermore, this error explosion happens for a smaller value of $C$ for the test set error than for the cross-validation error (i.e. the blue test curve is to the left of the red cross validation curve). In the right panel, with the SVM centroid intercept, the error rate does not explode; moreover, the test error curve behaves similarly to the cross-validation curve. The curves on the right and left panels are identical for $C > 10^{-2}$. For this data set, 5-fold cross-validation gives a test set error of 28.1% for the regular SVM intercept, but 24.35% for the SVM centroid intercept.

Over 200 repetitions of this simulation, regular SVM has an mean test set error of 25.95% (MD gives 23.95%). If we replace the regular SVM intercept, $b_{svm}$ with $b$ defined above we get an average test set error of 24.80%; this intercept gives an average improvement of 1.15% for this dataset (this difference is statistically significant using a paired t-test which gives a p-value of $2 \times 10^{-16}$).

When the classes are very unbalanced other error metrics are used (e.g. F-score, AUC, Choen's Kappa, etc). If AUC is used i.e. the intercept is tuned independently of the direction, issues with the intercept discussed in this section will not occur. However, when other metrics are used the improved intercepts will likely be more effective.

The intercept $b$ defined above will not improve SVM's performance in all scenarios, but is not likely to harm the performance. The intercept $b$, however, is simple to implement and can give a better test set error.

## 3.7 Discussion

### 3.7.1 Geometry of Complete Data Piling

Theorems 3.3.2 and 3.3.4 give further insight into the geometry of complete data piling directions. In this section we consider directions to be points on the unit sphere; the equivalence class of a single direction is represented by two antipodal points.

When $d \geq n$ there are an infinite number of directions $P$ that give complete data piling. If we restrict ourselves to the $n$ dimensional subspace generated by the data there are still an infinite number of di-

rections that give complete data piling (Ahn and Marron, 2010b); within this subspace $P$ forms a great circle of directions. Theorem 3.3.2 says that if we further restrict ourselves to the $n-1$ dimensional affine hull of the data there is only a single direction of complete data piling and this direction is the maximal data piling direction. The aforementioned great circle of directions intersects the subspace parallel to the affine hull of the data at two points (i.e. a single direction).

Note Equation 3.2.3 shows $\mathbf{w}_{mdp}$ is a linear combination of the data and Theorem 3.3.2 shows furthermore that $\mathbf{w}_{mdp}$ an affine direction. Finally, Theorem 3.3.4 also characterizes the stronger condition when the MDP is a convex classifier (see Section 3.2.1) i.e. when the MDP direction points between the convex hulls of the two classes ($\mathbf{w}_{mdp} \in C$).

### 3.7.2 nu-SVM and the Reduced Convex Hull

A number of papers look at an alternative formulation of the SVM optimization problem (so called nu-SVM). These papers give an interesting, geometric perspective that characterizes soft margin SVM in terms of hard margin SVM (see citations in Section 3.1.2).

Recall the convex hull of a set of points is given by $H(\{\mathbf{x}_i\}_{i=1}^n) := \left\{\sum_{i=1}^m \lambda_i \mathbf{x}_i | \sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0\right\}$. Suppose we decrease the upper bound on the coefficients such that $\lambda_i \leq c$ for some $c \geq 0$. Define the *reduced convex hull* (RCH) as

$$R_c(\{\mathbf{x}_i\}_{i=1}^n) := \left\{ \sum_{i=1}^n \lambda_i \mathbf{x}_i | \sum_{i=1}^n \lambda_i = 1, \lambda_i \leq c \right\}$$

Note $R_c \subseteq H$, $R_c = H \Leftrightarrow c = 1$ and $c = \frac{1}{n} \Leftrightarrow R_c = \{\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\}$ (i.e. a single point). Also note that, $R_c$ is not necessarily a dilation of $H$ e.g. see Figure 5 from (Bennett and Bredensteiner, 2000) for an example. Furthermore, define $E_c$ to be the set of *extreme points* of $R_c$ (the RCH of a finite set of points is a polytope and the extreme points are the vertices of this polytope).

Similarly to Definition 3.2.1 of the convex directions for two classes, we define the set of *reduced convex directions*, $RC_c$

**Definition 3.7.1.** *Let* $0 \leq c \leq \min\left(\frac{1}{n_+}, \frac{1}{n_-}\right)$ *and let* $RC_c$ *denote the set of all vectors associated with the directions that go between the c reduced convex hulls convex hulls of the two classes i.e.*

$$RC_c = \{a(\mathbf{c}_+ - \mathbf{c}_-) | a \in \mathbb{R}, a \neq 0, \text{ and } \mathbf{c}_j \in R_c(\{\mathbf{x}_i\}_{i \in I_j}), j = \pm\}.$$

Similarly, let $ERC_c$ denote the set of extreme points of $RC_c$ (where the points are marked by their respective class labels). Note that even if the convex hulls of the two classes intersect, there (usually[2]) exists a $c' \geq 0$ such that the $c'$ reduced convex hulls of the two classes do not intersect.

The nu-SVM literature shows that for every $C$, there exists a $c \geq 0$ that soft margin SVM direction with tuning parameter $C$ is equivalent to the hard margin SVM direction of the extreme points of the $c-$reduce convex hull of the data ($ERC_c$) which are a subset of the convex hull of the original data.

We point this geometric insight out because it gives similar geometric insights into SVM as our paper. Furthermore, the RCH formulation connects soft margin SVM to the maximal data piling direction; in particular, soft margin SVM is the MDP of the extreme points of the RCH.

### 3.7.3 Relations Between SVM and Other Classifiers

We have shown SVM can be exactly or approximately equivalent to the mean difference or maximal data piling direction (or possibly cropped versions of these two classifiers). When the data are balanced and $C$ is sufficiently small, SVM becomes exactly the mean difference. When the data are unbalanced, SVM becomes a cropped version of the mean difference. Hard margin SVM is always the maximal data piling direction of the support vectors meaning it can be viewed as a cropped MDP. We gave conditions for when hard margin SVM is exactly the MDP of the full dataset.

These results are mathematically interesting i.e. they give conditions when a quadratic optimization problem reduces (exactly or approximately) to a problem which has a closed form solution with a simple geometric interpretation. By carefully studying how this behavior depends on the tuning parameter we give a number of insights into tuning SVM (see Section 3.6).

Furthermore, these insights can be directly relevant to the data analyst. For example, the analyst may learn something about the data when they encounter scenarios in which SVM is either exactly or approximately equivalent to one of these simple classifiers. In scientific applications using SVM, the data analyst may want to know more about why cross-validation selects a given tuning parameter.

Our results help both practitioners and researchers transfer intuition from the MD and MDP classifiers to SVM and vice versa. The mean difference classifier is widely used (especially if one takes the data transformation perspective from Section 3.2.2) and a lot is known about when it works well and doesn't

[2] If, for example, the class means are identical the RCH formulation may breakdown.

(e.g. if the two classes are homoskedastic point clouds). While the MDP is an active topic of research, as discussed in (Miao, 2015), we understand some cases when the MDP works well and doesn't.

Finally, the results in this chapter raise the question: how much performance gain does SVM achieve over more simple classifiers? For example, for a particular application it could be the case that the mean difference plus some combination of simple data transformation, robust mean estimation, and/or kernels would achieve a very similar test set error rate as SVM. This question is important to practitioners because more simple models are often favored for reasons of interpretability, computation, robustness, etc.

An interesting follow up question for researchers is to empirically compare SVM to a variety of mean difference and maximal data piling like classifiers for a large number of datasets. We suspect that in some cases, the more simple classifiers will perform very similarly to SVM and in other cases SVM will truly beat out these more simple classifiers. Finally, we recommend that practitioners keep track of at least the MD (and possibly MDP in high dimensions) when fitting SVM.

**Appendix A.**

In this section we prove Theorem 3.3.2. Online supplementary material including code to reproduce the figures in this chapter, proofs that were omitted for brevity and simulations can be found at: https://github.com/idc9/svm_geometry.

*Proof.* **of Theorem 3.3.2**

We first prove the existence and uniqueness of complete data piling directions $P$ in the affine hull of the data. We then show that this unique, affine data piling direction is in fact the direction of maximal data piling.

Recall we assume that $d \geq n-1$ and the data are in general position. Let the set of affine directions $A$ be given as follows

$$A = \{\mathbf{a}_1 - \mathbf{a}_2 | \mathbf{a}_j \in \text{aff}(\{\mathbf{x}_i\}_1^n), j = 1, 2\}.$$

Note that $A$ is the $n-1$ dimensional subspace parallel to the affine space $\text{aff}(\{\mathbf{x}_i\}_1^n)$ generated by the data i.e. $A$ contains the origin.

We first show that without loss of generality $d = n-1$. Note that both $A$ and $P$ are invariant to a fixed translation of the data. Therefore, we may translate the data so that $0 \in \text{aff}(\{\mathbf{x}_i\}_1^n)$ (e.g. translate by the

mean of the data). The data now span an $n-1$ dimensional subspace since the affine hull of the data now contains the origin. Furthermore, $\text{span}(\{\mathbf{x}\}_1^n) = \text{aff}(\{\mathbf{x}\}_1^n) = A$. Thus without loss of generality we may consider the data to in fact be $n-1$ dimensional (i.e. $d = n-1$).

We are now looking for a vector $\mathbf{v} \in A$ that gives complete data piling. Note by the above discussion and assumption we have $A = \mathbb{R}^d$. This means we are looking for $\mathbf{v} \in \mathbb{R}^d$ and $a, b \in \mathbb{R}$ with $a \neq 0$ satisfying the following $n$ linear equations

$$\mathbf{x}_i^T \mathbf{v} = a y_i + b \text{ for } i = 1, \ldots, n.$$

Since the magnitude of $\mathbf{v}$ is arbitrary we fix $a = 1$ without loss of generality. We now have

$$\mathbf{x}_i^T \mathbf{v} = y_i + b \text{ for } i = 1, \ldots, n$$

which can be written in matrix form as

$$X\mathbf{v} + b\mathbf{1}_n = \mathbf{y} \tag{3.7.1}$$

where $X \in \mathbb{R}^{n \times d}$ is the data matrix whose rows are the data vectors $\mathbf{x}_i$ and $\mathbf{y} \in \mathbb{R}^n$ is the vector of class labels. This is a system of $n$ equations in $\mathbf{R}^{d+1}$ which can be seen by appending 1 onto the end of each $\mathbf{x}_i$ i.e. $\tilde{\mathbf{x}}_i = (\mathbf{x}_i, 1) \in \mathbb{R}^{d+1}$ and letting $\mathbf{w} = (\mathbf{v}, b)$. Then Equation 3.7.1 becomes

$$\widetilde{X}\mathbf{w} = \mathbf{y} \tag{3.7.2}$$

where $\widetilde{X} \in \mathbb{R}^{n \times d+1}$ is the appended data matrix.

Recall that we assumed $d = n-1$ so Equation 3.7.2 is a system of $n$ equations in $\mathbb{R}^n$. Further recall that the data are in general position meaning that the $n$ data points are affine independent in the $n-1$ dimensional subspace of the data. Affine independence is equivalent to linear independence of $\{(\mathbf{x}_i, 1)\}_1^n$. Therefore the matrix $\widetilde{X} \in \mathbb{R}^{n \times n}$ has full rank and Equation 3.7.2 always has a solution, $\mathbf{v}^*$, and this solution is unique.

Existence of a solution to Equation 3.7.2 shows that $P \cap A \neq \emptyset$. Uniqueness of the solution to Equation 3.7.2 shows that this intersection $P \cap A$ can have only one direction of which $\mathbf{v}^*$ is a representative element.

We now show that $\mathbf{v}^*$ is in fact the maximal data piling direction. We no longer assume that $d = n-1$.

We first construct an orthonormal basis $\{\mathbf{t}_i\}_1^d$ of $\mathbb{R}^d$as follows. Let the first $n-1$ basis vectors $\mathbf{t}_1,\ldots,\mathbf{t}_{n-1}$ span $A$. Let $\mathbf{t}_n$ be orthogonal to $A$ but in the span of the data $\{\mathbf{x}_i\}_1^n$ (recall the data span an $n$ dimensional space while the affine hull of the data is $n-1$ dimensional). Let the remaining $d-n+1$ basis vectors be orthogonal to $A$ and the span of the data.

We show that the vector $\mathbf{t}_n$ projects every data point onto a single point i.e. $\mathbf{x}_i^T\mathbf{t}_n = c$ for each $i = 1,\ldots n$ and some $c \in \mathbb{R}$. Suppose we translate aff$(\{\mathbf{x}_i\}_1^n)$ along $\mathbf{t}_n$ until the origin lies in the affine hull of the translated data. In particular, the data now span an $n-1$ dimensional subspace that is orthogonal to $\mathbf{t}_n$ (where as before they spanned an $n$ dimensional subspace). We now have that for some $c \in \mathbb{R}$

$$\mathbf{t}_n^T(\mathbf{x}_i + c\mathbf{t}_n) = 0 \text{ for each } i = 1,\ldots,n$$

$$\mathbf{t}_n^T\mathbf{x}_i = c \text{ for each } i = 1,\ldots,n$$

since $\mathbf{t}_n$ is unit norm.

Let $\mathbf{v} \in \mathbb{R}^d$ be a representative vector of the direction in the affine hull of the data that gives complete data piling (given above). Suppose $\mathbf{v}$ has unit norm and is oriented such that

$$\mathbf{v}^T\mathbf{x}_i = ay_i + b$$

for some $a, b \in \mathbb{R}$ with $a > 0$ (note fixing $a > 0$ eliminates the antipodal symmetry of data piling vectors).

We now show that $\mathbf{v}$ is in fact the maximal data piling direction. Let $\mathbf{w} \in \mathbb{R}^d$ be another vector with unit norm that gives complete data piling (i.e. $\mathbf{w} \in P$). In particular, there exists $a_v, a_w, b_v, b_w \in \mathbb{R}$ with $a_v, a_w > 0$ such that

$$\mathbf{v}^T\mathbf{x}_i = a_v y_i + b_v \text{ for each } i = 1,\ldots,n.$$

$$\mathbf{w}^T\mathbf{x}_i = a_w y_i + b_w \text{ for each } i = 1,\ldots,n.$$

Assume for the sake of contradiction that $\mathbf{w}$ projects the data possibly further apart than $\mathbf{v}$ does. In particular assume that $a_w \geq a_v$.

Since $\{\mathbf{t}_i\}_1^d$ is a basis we can write

$$\mathbf{w} = \sum_{i=1}^{d} \alpha_i \mathbf{t}_i.$$

Next compute the dot products with the data. For any $j = 1, \ldots, n$,

$$\mathbf{w}^T \mathbf{x}_j = \left( \sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i \right)^T \mathbf{x}_j + \alpha_n \mathbf{t}_n^T \mathbf{x}_j + \sum_{i=n+1}^{d} \alpha_i \mathbf{t}_i^T \mathbf{x}_j.$$

Recall the basis vectors $\mathbf{t}_{n+1}, \ldots, \mathbf{t}_d$ are orthogonal to the data points so the third term in the sum is zero. Furthermore, the dot product of $\mathbf{t}_n$ with each data point is a constant. Thus we now have

$$\mathbf{w}^T \mathbf{x}_j = \left( \sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i \right)^T \mathbf{x}_j + \alpha_n c, \text{ for all } j = 1, \ldots, n.$$

Thus we can see the vector

$$\mathbf{w}' = \sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i$$

also gives complete data piling. However this vector lies in $A$ since it is a linear combination of the first $n-1$ basis vectors. We have shown that there is only one direction in $A$ with complete data piling thus $\sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i \propto \mathbf{v}$. In particular, for some $\alpha > 0$

$$\sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i = \alpha \mathbf{v}.$$

So we now have

$$\mathbf{w}' = \alpha \mathbf{v} + \alpha_n \mathbf{t}_n.$$

Recall $||\mathbf{v}|| = ||\mathbf{w}|| = 1$ and $\mathbf{t}_n$ is orthogonal to $\mathbf{v}$ by construction. Therefore $\alpha^2 + \alpha_n^2 = 1$. In particular if $\alpha_n > 0$ then $\alpha < 1$.

Let $\mathbf{x}_+$ and $\mathbf{x}_-$ be any point from the positive and negative class respectively. By construction we have

$$\mathbf{v}^T (\mathbf{x}_+ - \mathbf{x}_-) = a_v.$$

$$\mathbf{w}^T (\mathbf{x}_+ - \mathbf{x}_-) = a_w.$$

However expanding this last line we get

$$\mathbf{w}^T (\mathbf{x}_+ - \mathbf{x}_-) = (\alpha \mathbf{v} + \alpha_n \mathbf{t}_n)^T (\mathbf{x}_+ - \mathbf{x}_-)$$

$$\mathbf{w}^T(\mathbf{x}_+ - \mathbf{x}_-) = \alpha \mathbf{v}^T(\mathbf{x}_+ - \mathbf{x}_-) + \alpha_n \mathbf{t}_n^T(\mathbf{x}_+ - \mathbf{x}_-).$$

But $\mathbf{t}_n^T\mathbf{x}_+ = \mathbf{t}_n^T\mathbf{x}_- = c$ so the last term is zero. Thus we now have

$$\mathbf{w}^T(\mathbf{x}_+ - \mathbf{x}_-) = \alpha a_v.$$

Thus

$$\alpha a_v = a_w.$$

However unless $\mathbf{w} = \mathbf{v}$ (so $\alpha_n = 0$) we have $0 < \alpha < 1$. Therefore $a_w < a_v$ contradicting the assumption that $a_w \geq a_v$. Therefore $\mathbf{v}$ is the maximal data piling direction. ■

**Appendix B.**

Theorem 3.3.4 gives a geometric characterization when the set of convex directions intersects the set of complete data piling directions. We can also characterize this event through a linear program.

An alternative way of deciding if $C \cap P = \emptyset$ and computing the intersection if it exists is through the following linear program (proof of Theorem 3.7.2 is a straightforward exercise in linear programming).

**Theorem 3.7.2.** *$C \cap P \neq \emptyset$ if and only if there is a solution to the following linear program*

$$\begin{aligned}
&\underset{\alpha\in\mathbb{R}^{n_+},\beta\in\mathbb{R}^{n_-},\mathbf{v}\in\mathbb{R}^d,b\in\mathbb{R}}{\textit{minimize}} && 1 \\
&\textit{subject to} && X\mathbf{v} + \mathbf{1}_n b = \mathbf{y} \\
& && \sum_{i\in I_+} \alpha_i \mathbf{x}_i - \sum_{i\in I_-} \beta_i \mathbf{x}_i = \mathbf{v} \\
& && \sum_{i\in I_+} \alpha_i = 1 \\
& && \sum_{i\in I_-} \beta_i = 1 \\
& && \alpha_i, \beta_i \geq 0 \textit{ for } i = 1,\ldots,n.
\end{aligned} \tag{3.7.3}$$

*In the case a solution* $\mathbf{v}$ *exists then* $\mathbf{v} \in C \cap P$.

The vector $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of ones, $X$ is the $\mathbb{R}^{n\times d}$ data matrix and $\mathbf{y} \in \mathbb{R}^n$ is the vector of class labels. The first constraint says $\mathbf{v}$ must be a complete data piling direction, $\mathbf{v} \in P$. The remaining constraints say $\mathbf{v}$ must be a convex direction, $\mathbf{v} \in C$.

Note that solving this linear program is at least as hard as solving the original SVM quadratic program therefore Theorem 3.7.2 is not of immediate computational interest. This theorem, however, does give an alternate mathematical description $C \cap P \neq \emptyset$ which may be of theoretical interest.

CHAPTER 4

# Vertex Centrality Metrics

A *network* is a collection of *vertices* (e.g. Supreme Court opinions) and *edges* between them (e.g. citations). A major question in network analysis is to identify *important* vertices within a network. A *vertex centrality metric* (Kolaczyk, 2009) assigns a number to each node in a network based on how important that node is *in some sense*. For example, the field of academic ego stroking, scholars count the number of citations their papers receive[1]. Another example comes from *information retrieval* where web pages are ranked by how "authoritative" they are based on which other webpages link to them[2]. The *PageRank* algorithm is a vertex centrality metric which is (was?) famously an important part of Google's initial success as a search engine (Bryan and Leise, 2006).

While there are a large number of vertex centrality metrics (e.g. in-degree, eigenvector centrality, betweenness centrality, PageRank, etc), there are few empirical or theoretical tools available to understand which vertex centrality metric is right for a given network. In (Carmichael et al., 2017), reproduced in Chapter 5, we develop a methodology to empirically evaluate vertex centrality metrics for an evolving network. The idea is to measure how predictive a given centrality metric is of future citations (e.g. if a case has a large PageRank this year, it's likely to be cited next year). The results of applying this methodology to the US Supreme Court citation network are discussed in (Carmichael et al., 2017). In this chapter we give a brief overview of the methodology and discuss some future research directions.

## 4.1 Networks background

Formally we write a network $G = (V, E)$ were $V$ is the set of vertices and $E$ is the set of edges. We will write $V$ for both the set of vertices and its cardinality (i.e. number of vertices). Similarly for $E$. Edges may be either directed or undirected (we do not consider self edges or multi-edges).

[1] In this case the nodes of the network are papers and the (directed) edges correspond to citations from one paper to another. The number of citations a paper has received is then its *in-degree*.

[2] In this case the nodes of the network are webpages and edges are hyperlinks between webpages.

We consider a network which evolves over time such that new nodes and edges are added sequentially, but nodes/edges cannot be removed. A citation network is an example of such a network. In this case, we write $G_1 = (V_1, E_1), G_2 = (V_2, E_2), \ldots, G_T = (V_T, E_T)$ for the network at times $t_1 \leq t_s \leq \cdots \leq T$ (note $V_{t_i} \subseteq V_{t_j}$, $E_{t_i} \subseteq E_{t_j}$ for $i < j$). We assume all of a node's edges are added at the time the node enters the system.

A vertex centrality metric is a measure of how "important" each vertex is in a network. A metric, $m : V to \mathbb{R}^V$, assigns a number to each vertex with the convention that larger values of $m$ mean "more important". In a directed network, the most simple centrality metric is the *in-degree* i.e. the number of edges pointing to that node. Another class of centrality metrics, called *eigenvector centrality metrics*, is based on the notion that a node is important if it has been linked to a lot by nodes who are themselves important. Examples of such metrics include: *eigenvector centrality*[3], PageRank, Katz centrality, hubs and authorities (Kolaczyk, 2009). Another set of metrics is based on the notion that important vertices are "centrally located" in the graph. For example, for a given vertex, $v$, *betweenness centrality* counts the number of times $v$ lies on the shortest path between two other nodes in the graph. For a modern review of vertex centrality metrics see (Boldi and Vigna, 2014).

### 4.2 Sort experiment methodology

The goal of this section is to develop an empirical method to evaluate/compare different vertex centrality metics in an evolving (citation) network. The core assumption is that a better vertex centrality metric will better predict future citations. To quantify how predictive a given centrality metric is we measure it's performance as a *link prediction* (Liben-Nowell and Kleinberg, 2007) method by performing the following computation which we refer to as the "sort experiment."

For expositional simplicity we assume that at every time step one node enters the system[4] i.e. at time $t_n$ the network, $G_n = (V_n, E_n)$ has $n$ nodes. At time $t_{n+1}$ the $n+1$st node and all its edges are added. Fix a vertex centrality metric $m$ and order the nodes from largest to smallest values of their metrics (recall a large value of $m$ means a "better" vertex). Let $v^m_{(1),n}, v^m_{(2),n}, \ldots, v^m_{(n),n}$ be the vertices ordered by $m$ at time $n$ (here $(k)$ means the $k$th largest value).

[3] Confusingly, this term refers to both a whole family of metrics and one particular member of that family. No one has ever accused the mathematical sciences of employing good naming conventions.

[4] Relaxing this assumption is straightforward.

Suppose the $n+1$st node enters the network and attaches to $K$ existing nodes. Using the ranking determined by $m$, suppose the nodes cited by this new vertex are $v^m_{(r_1),n}, v^m_{(r_2),n}, \ldots, v^m_{(r_K),n}$ where $r_1, \ldots, r_K$ are the ranks of these cited nodes. The *mean rank score* (Zanin et al., 2009) is given by

$$MRS^m_n := \frac{1}{K}\sum_{i=1}^{K}\frac{r_i}{n} \tag{4.2.1}$$

We exclude vertices which do not cite anyone.

**Remark 4.2.1.** If the metric $m$ is very predictive of future citations, the ranks will tend to be small therefore the smaller the value of mean rank score, the better the metric.

For a network with $V$ total vertices we define the average mean rank score as

$$AMRS^m := \text{average}\left\{MRS_n \text{ s.t. vertex } n+1 \text{ links to at least one existing vertex}, 2 \leq n \leq V-1\right\} \tag{4.2.2}$$

To evaluate and compare a set of vertex centrality metrics, $\{m_1, \ldots, m_N\}$ we compute the average rank score for each metric.

**Remark 4.2.2.** Computing $AMRS^m$ requires computing the vertex centrality metric $m$ for every time step in the network. This means $AMRS^m$ may be computationally prohibitive e.g. if $m$ takes $O(V)$ time to compute then (naively) $AMRS^m$ would take $O(V^2)$ time to compute[5].

We therefore approximate $AMRS^m$ by computing a smaller, random subset of the mean rank scores. In particular, we compute the metric $m$ at $S$ (e.g $S = 100$) snapshots of the network a times $t_1, \ldots, t_S$. We use these the centrality metrics calculated at these snapshots to approximate the centrality metrics for some short time interval where (hopefully), the value of $m$ does not vary too much.

### 4.2.1 Discussion

There are potential alternatives to the *sort experiment* described in Section 4.2. We could cast the problem as a binary classification (i.e. link prediction) problem where, at each time step, the outcome, $y$ is whether or not an existing vertex, $v$, will be cited by a new vertex and the $X$ data is the value of $v$'s vertex centrality metric. One could then train a model such as logistics regression or support vector machine then evaluate the model on a test set. There are several issues with this classification approach. For

[5]For some metrics such as in-degree, an online algorithm might go down to $O(V)$ time complexity

example, this is a very unbalanced dataset (i.e. most $y$ values are “not cited") and low information (i.e. the value of the vertex centrality metric is a very noisy predictor for a citation). Most measures of classification accuracy on a test set (e.g. misclassification error or the logistic loss) will likely be noisy. Moreover, any linear classifier would almost certainly put a positive value on the coefficient of the vertex centrality metric and therefore the ranks predictions would be equivalent to the ranks from the sort experiment. Thus the sort experiment can be viewed as a non-parametric version of fitting such a statistical model.

In the field of information retrieval there are a variety of metrics used to quantify how well a given ranking algorithm is performing such as precision, recall, precision at K, and reciprocal rank (Murphy, 2012). These metrics are typically used to evaluate search engines where one expects the selected results to be near the top of the list. However, for “sort experiment" based on the vertex centrality metics along, we would not expect the cited cases to be near the top of the list. We do, however, expect a vertex metric captures some signal which we believe makes the mean rank score the most appropriate metric. This assumption, however, requires further investigation.

There are a number of other theoretical questions which we leave for future work. For example, we want to be able to compare the value of the average mean rank score for two different metrics (i.e. how much lower is meaningful). To do this we need something like a confidence interval, however, it's not totally clear what statistical model one should assume to construct a confidence interval. One relevant class of models is a generalization of *preferential attachment*.

Recall *preferential attachment* model is simple probabilistic model for growing a network (Barabási and Albert, 1999; Van Der Hofstad, 2016). In the simplest case, a network is grown by successively adding vertices. When a new vertex, $n$, enters the system it randomly selects an existing vertex, $v$, to attach to with probability proportional to $v$'s current degree. Another model, called *preferential attachment with fitness*, has similar dynamics, but in this case each vertex has a random *fitness value* and is selected with probability proportional to it's in-degree plus its fitness value (Borgs et al., 2007). Many other generalizations abound.

Consider the following generalization of preferential attachment. Pick a vertex centrality metric, $m$ (e.g. PageRank). When a new node, $n$, enters the system it attaches to an existing vertex, $v$ with probability proportional to $v$'s current value of $m$. We call this $m$-preferential attachment. We can similarly define $m$-preferential attachment with fitness.

A natural direction is to study the “sort experiment" under different $m$-preferential attachment models. For example, one would expect that on a network grown according to $m$-preferential attachment the $m$ metric would perform better on the “sort experiment" than another metric $\tilde{m}$. Things might get more interesting when considering the fitness version of $m$-preferential attachment (i.e. the performance of the $m$ metrics might be more comparable to other metrics).

CHAPTER 5

# Vertex Centrality Metrics with Applications to the Law

## 5.1 Introduction

Precedent– the rule that judges rely on past decisions in adjudicating the disputes in front of them——is a fundamental part of the American legal system. Precedent plays a role in understanding how the law evolves, determining how the courts behave, and helping legal search engines recommend relevant cases (Landes and Posner, 1976). Studying the network of citations between judicial opinions can provide an empirical perspective on precedent. This Comment examines the Supreme Court citation network and the Federal Appellate citation network.[1] Building upon previous empirical legal work, this Comment uses network analysis to provide insights into the role of precedent in the courts and to identify important cases in various contexts. Vertex centrality metrics, which measure how important a vertex is in a network in different ways, provide a way of quantifying the notion of importance of a case in a citation network. There are many kinds of vertex centrality metrics. This Comment further develops a methodology to evaluate vertex centrality metrics in an evolving network based on how predictive a metric is of future citations. This methodology is able to identify several possibly surprising results regarding court behavior and behavior of the metrics themselves. In particular, it unexpectedly shows that the number of cases cited in an opinion is a stronger predictor of whether that opinion will be cited in the future than the number of times that opinion has already been cited by other opinions.

The broader aim of this research is to understand the factors driving the evolution of the law. The law evolves incrementally by building on precedent as judges answer novel questions based on principles set out in prior cases(Landes and Posner, 1976). Understanding the precedential weight of a case is a challenging but productive task for several reasons. Scholars, for instance, use precedent to examine what factors influence the evolution of the law, identify which issues are currently most relevant, and predict what issues might become active in the future (Cross and Spriggs, 2010). Practitioners are often

[1] The citation network here means the network of cases and the citations between them.

required to identify relevant cases that are most likely to convince a judge (Bennardo, 2014). Additionally, nonprofit organizations, researchers, and companies might want to build legal research tools.[2]

One way that precedent can be examined is through citations in written judicial opinions. This method of analysis operates on the assumption that "[e]ach judicial citation contained in an opinion is essentially a latent judgment about the case cited." (Fowler and Jeon, 2008)[3] In other words, a citation is a good indication that a cited case is precedent for the case at hand (Fowler and Jeon, 2008). Based on this assumption, scholars have begun using empirical methods based on the network of legal citations to rank the value of cases (Fowler et al., 2007).

Formally, a network is a collection of objects (called vertices or nodes) and connections between them (called edges) (Kolaczyk, 2009). The study of networks has become popular in recent decades the internet is a network of computers, Facebook and Twitter capture human social networks, (Leskovec and Mcauley, 2012) and neuroscientists study the brain as a network of neurons connected by white matter fiber tracts (Craddock et al., 2013). Another popular subject of study is the citation networks of academic papers (Leskovec et al., 2005). This Comment examines legal citation networks, defined as networks of judicial opinions and the citations going between them.[4] Each case is a vertex, and each citation is an edge. Figure 5.1.1 shows an example of a network. In particular, the vertices (dots) represent Roe v. Wade (Rowe, 1973) and cases that either cite to or are cited by Roe v. Wade. The edges (lines) represent citations between this set of cases.

There are many different ways to quantify the importance of a vertex in a network, called vertex centrality metrics. Two of the simplest vertex centrality metrics are in-degree and out-degree. In-degree is the count of citations a case has received, while out-degree is the count of cases cited in an opinion. Citations are directed backwards in time, which means that a citation would go out from a case in 2017 and in to a case in 1990. Figure 5.1.2 illustrates a simple example of a hypothetical citation network with six nodes/cases (dots) and eight directed edges/citations (arrows). The highlighted case A is cited by

[2]For an example of a legal research tool and how they may operate, see Natural Language Searches, Thomson Reuters Westlaw https://lawschool.westlaw.com/marketing/display/RE/151

[3]Some scholars have resisted this context-neutral approach to citations and prefer to use data from LexisNexis's citator tool, Shepard's, or Westlaw's citator tool, KeyCite, to only capture citations with a positive valence. See, e.g. (Hitt, 2016)

[4]For an example of a legal citation network that includes statutory provisions, see (Bommarito, 2010).

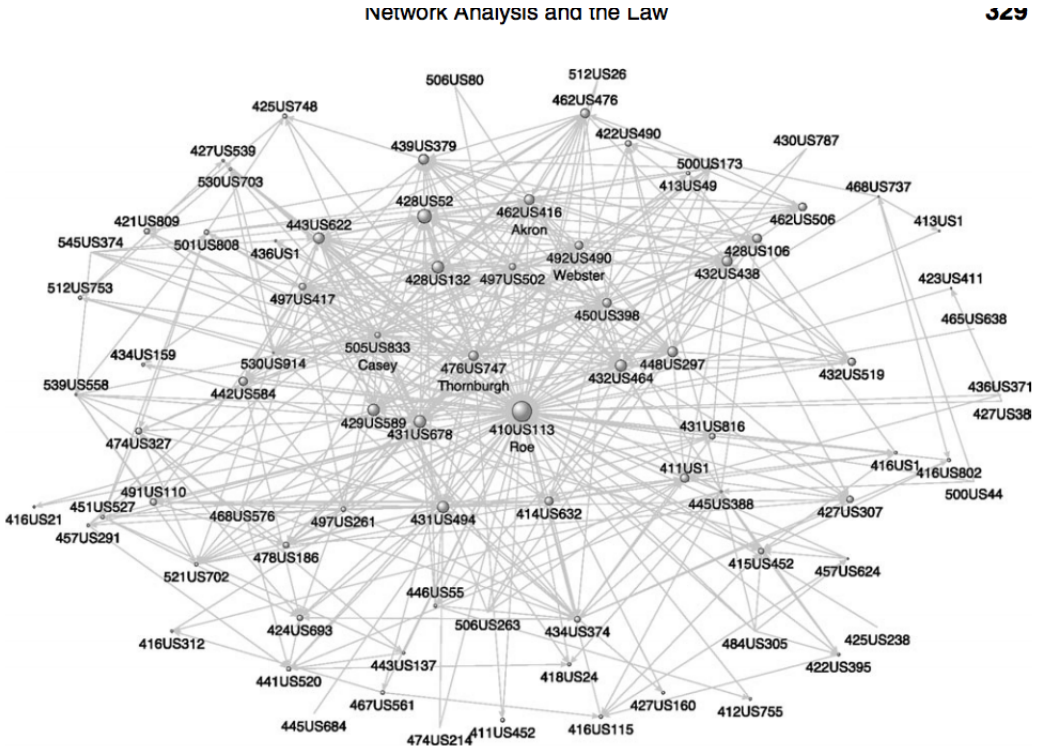



**Fig. 2** Extended network of abortion decisions.

**Figure 5.1.1:** A visual of a network; this plot shows Roe v. Wade and its neighboring cases. Each dot represents a case and each gray arrow represents a citation from one case to another. This graphic is reproduced from (Fowler et al., 2007).

three other cases meaning it has an in-degree of three. Similarly, case A cites two other cases meaning it has an out-degree of two.

In addition to the in-degree and out-degree metrics, there are many more sophisticated measures of vertex centrality that are built on other assumptions about how the structure of a network reflects importance. However, despite the large number of vertex centrality metrics, there is little empirical work that evaluates which centrality metric is the right choice for a given task.

This Comment develops a novel methodology to evaluate vertex centrality metrics in a citation network. The core assumption of this methodology is that a good vertex centrality metric should be able to predict future citations. Because of the theoretical foundation established by legal scholars connecting

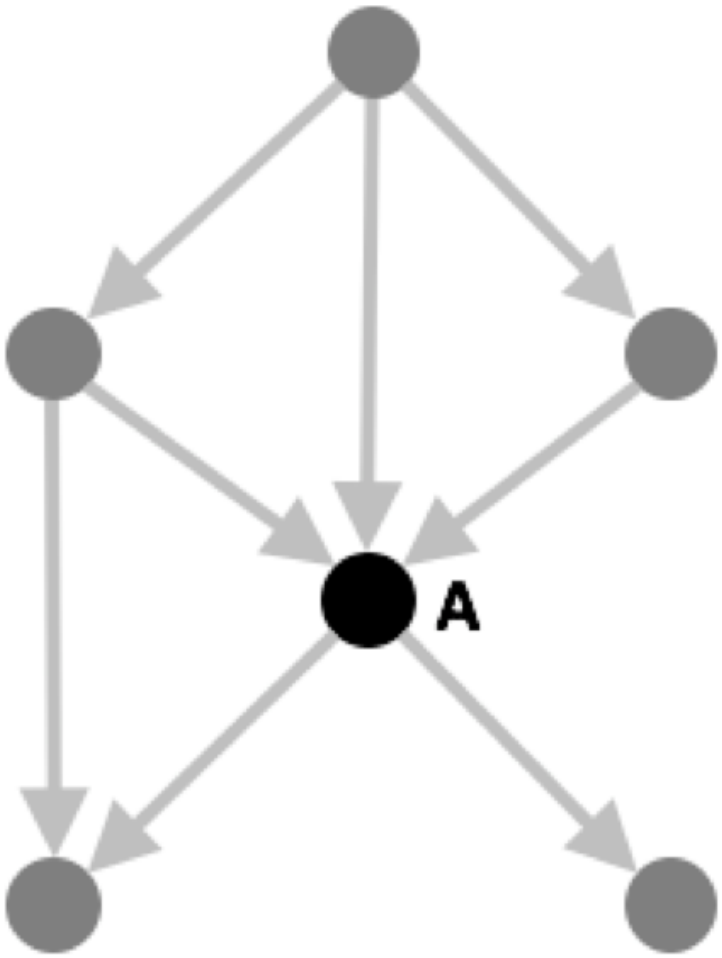



**Figure 5.1.2:** A simple example of a citation network with six cases. The highlighted case A has been cited three times. Therefore, case A has an in-degree equal to three. Similarly, case A cites two cases and therefore has an out-degree equal to two.

citations to precedent, the results of this methodology may have important implications for the study of precedent.[5]

Furthermore, this Comment models methods of reasoning that lawyers will have to engage in more frequently in the future. The number of areas of legal practice in which lawyers would benefit from a baseline understanding of the principles underlying legal tools is growing (Henderson, 2015). Therefore lawyers will have to make informed decisions about technical statistical issues. Legal research, electronic discovery, and transactional practice are all increasingly being shaped by complicated statistical

[5]This is not to say that legal citation network analysis is a priori an analysis of precedent, nor is it to say that precedent is the only area of legal study that could benefit from network analysis.

methods, (Lohr, 2017) and statistical methods are creating new categories of tools and services, such as outcome prediction in litigation.[6] Thus, lawyers in a growing number of fields stand to gain from understanding statistical concepts and increasingly risk liability if they do not make themselves aware of how their tools work.[7]

This Comment further develops the use of citation network analysis to study legal precedent. Our findings are based on a novel statistical methodology, which we developed to empirically compare vertex centrality metrics in a citation network. We find novel evidence to support previous claims made about the nature of precedent and argue that how well-grounded an opinion is in existing precedent, the presence of multiple opinions in a case, and the age of an opinion all contribute to how likely a case is to be cited in the future.

This Comment proceeds in four parts. Part 5.2 surveys prior relevant work on American case law citation networks. Part 5.3 gives additional background information on vertex centrality metrics and presents a detailed description of the data and the methodology. Part 5.4 presents the results of the statistical analyses of both the United States Supreme Court network and the Federal Appellate network. The methodology demonstrates that out-degree is more predictive of future citations than in-degree, which was an unexpected result. Less surprisingly, the methodology also demonstrates strong aging effects in the network, or that newer cases are more likely to be cited than older cases. The results reaffirm and build upon existing research in both assessments of vertex centrality metrics and of the effect of time on citation rates. Part 5.5 discusses the legal and conceptual importance of these results. Specifically, it speculates about which qualities of cases will drive out-degree performance, particularly how well supported a case is and the presence of multiple opinions within a case. The presence of aging effects in the network is explained from both a legal and statistical perspective. Part 5.5 concludes by arguing that one vertex centrality metric, which did not perform well at predicting future citations, is biased towards cases of first impression.

---

[6] See, e.g., https://lexmachina.com/what-we-do/how-it-works/

[7] See, e.g., http://www.abajournal.com/news/article/lawyers_e_discovery_error_led_to_release_of_confidential_wells_fargo_client/

### 5.2 Precedent and case law citation network research

Landes and Posner, (Landes and Posner, 1976), define precedent as "something done in the past that is appealed to as a reason for doing the same thing again."[8] Precedent in case law develops based on analogical reasoning where one decision can "be an authority for another is that the facts are alike, or, if the facts are different, that the principle which governed the first case is applicable to the variant facts."(Landes and Posner, 1976) The principle that allows precedent to control is stare decisis. Stare decisis is "the doctrine of precedent, under which a court must follow earlier judicial decisions when the same points arise again in litigation."(Garner, 2014)

Precedent and stare decisis operate in a variety of ways[9]. The concept that a lower court must follow the decisions of a higher court in the same jurisdiction is referred to as "vertical" stare decisis (Gerhardt, 2011). "Horizontal" stare decisis, on the other hand, is the doctrine that a court, generally an appellate court, "must adhere to its own prior decisions, unless it finds compelling reasons to overrule itself." (Gerhardt, 2011) Other authorities are merely persuasive, meaning that they are "not binding on a court, but . . . [are] entitled to respect and careful consideration." [10]

Past empirical research on legal citation networks has been largely concerned with precedent. This research operates on the principle that "[e]ach judicial citation contained in an opinion is essentially a latent judgment about the case cited." (Fowler and Jeon, 2008) Or, rather, the fact that a judge has taken the time to include a citation to a particular case is a judgment on the quality of that case. Studying patterns of citations enables a better understanding about the evolution, growth, and state of the law.

Early empirical studies of precedent through citations involved counting and collating the citations a court made in a given time period (Merryman, 1953). Or, in network terms, early empirical studies of precedent relied on in-degree and out-degree. Scholars have used and continue to use this method to examine the writings of a range of courts, including the Supreme Court of California, (Merryman, 1953)

[8] Black's Law Dictionary defines precedent as "an action or official decision that can be used as support for later actions or decisions" or "a decided case that furnishes a basis for determining later cases involving similar facts or issues." (Garner, 2014)

[9] In addition to the descriptive discussion of precedent and stare decisis, there is an active normative discussion of the power of precedent, or rather, how much deference courts should give to past decisions (Gerhardt, 2011).

[10] Persuasive authorities include cases decided in a neighboring jurisdiction, which a court might evaluate "without being bound to decide the same way." (Landes and Posner, 1976)

federal circuit courts,(Landes and Posner, 1976) and the Supreme Court of the United States.[11] With the dramatic increase of computing power and storage of the last twenty years and the corresponding increased electronic availability of legal information, especially case law, scholars have been able to assess larger numbers of cases and citations (Landes and Posner, 1976; Merryman, 1953; Fowler and Jeon, 2008). At the same time, scholars have also begun using more sophisticated and computationally intensive network methods to measure the positions of cases and patterns of citation within bodies of law (Fowler et al., 2007). In fact, the Supreme Court citation network is used with some frequency to demonstrate novel vertex centrality measures and other network methods (Patty et al., 2013; Taylor et al., 2017). Moreover, it is used as an introduction to the subject of networks and vertex centrality in at least one introductory quantitative research book (Arnold and Tilton, 2015).

One common application of vertex centrality measures is to use them as a proxy for overall importance and, in turn, use them to rank cases. Perhaps the most forward example of this ranking genre is presented by Cross and Spriggs in their article, *The Most Important (and Best) Supreme Court Opinions and Justices* (Cross and Spriggs, 2010). The authors counted citations to Supreme Court opinions by the Supreme Court, the circuit courts, and the district courts and incorporated a more sophisticated network centrality into their rankings [12]. The authors found that a wide range of factors influence how strong a precedent is, including the issue area of the case, the age of the case, and the length of the opinion. Somewhat surprisingly, they found that unanimous decisions are less influential at the Supreme Court level, and that decisions by minimum-winning coalitions are no less influential than other decisions.

Another example of the ranking genre, and an early example of a study using a network centrality measure other than in-degree, is found in Fowler and Jeon's *The Authority of Supreme Court Precedent* (Fowler and Jeon, 2008). The authors analyzed the Supreme Court network using a more sophisticated vertex centrality metric and compared the results to lists of important Supreme Court cases compiled by legal experts and published by Congressional Quarterly, the Legal Information Institute, and the *Oxford Guide*.[13] Fowler and Jeon found that all ten of their most highly ranked authorities "are considered to be important by either Congressional Quarterly, the Legal Information Institute, or the *Oxford*

[11] A closely parallel, but conceptually distinct, line of study using the same method of citation counting is the study of the influence of individual judges (Landes et al., 1998).

[12] More specifically, the authors use a "legal relevance score," which is related to hubs and authorities. See Appendix 8.5 for a discussion of hubs and authorities.

[13] More specifically, the authors use hubs and authorities, which is discussed at more length in Appendix A.

*Guide*."(Fowler and Jeon, 2008) They further parsed their results by issue area and found similarly strong correlations between opinions of legal experts and the top five opinions identified as authorities within the broad categories of civil rights, criminal law, First Amendment, and privacy.

Other scholars employ these measures to study various qualities of case law (Black and Spriggs, 2013). One study examined the speed at which the probability of a case to be cited changed over time and found that a case's chances of being cited[14] "depreciate[s] about 81 percent and 85 percent between [its] first and 20th years of age at the Supreme Court and courts of appeals, respectively."(Neale, 2013) Multiple studies have examined differences in citation patterns across levels of the judicial hierarchy (Cross and Spriggs, 2010; Hitt, 2016). One demonstrated that the Supreme Court frequently cited "doctrinal paradoxes…, opinions of the Court for which every rationale for the Court's judgment is rejected by a majority," in an iterative process of working through complicated legal questions (Hitt, 2016). Conversely, doctrinal paradoxes were no more or less likely to be cited at the circuit court level and were less likely to be cited at the district court level (Hitt, 2016).

At a more sophisticated level, some papers have evaluated rankings produced by vertex centrality metrics. These papers evaluate vertex centrality metrics rankings based on a comparison to some external factor, such as expert opinion[15] or page views(Fowler and Jeon, 2008; Cross and Spriggs, 2010; Neale, 2013). One article, *Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court*, evaluates vertex metrics based on their ability to predict if a case will be cited in an upcoming year (Fowler et al., 2007). This Comment's methodology is similar to that of *Network Analysis and the Law*, but it uses richer data and examines potential citations at the case level as opposed to aggregating all cases in each year.

### 5.3 Methods

#### 5.3.1 Vertex Centrality Metrics

There are a number of different kinds of vertex centrality metrics. The most popular ones can be grouped into three categories: degree-based, eigenvector-based, and positional (Kolaczyk, 2009). Degree-based metrics include in-degree and out-degree; these metrics simply count raw numbers of

[14] Canadian cases, apart from those of the Supreme Court, are rarely cited more than 15 years after publication.

[15] Other papers have used multiple vertex centrality metrics without a method to evaluate those rankings.

citations.[16] Degree-based measures include in-degree and out-degree. While the various centrality metrics are often related, they are driven by different structural properties of the network.

The class of eigenvector centrality metrics is based on the idea that a case is important if it is cited by a lot of cases that are themselves important (Kolaczyk, 2009). Eigenvector centrality metrics judge a case to be more important if it is cited by many cases that are themselves cited by many other cases. Figure 5.3.1 demonstrates this idea in a small, hypothetical citation network; the circles represent cases and the arrows represent citations. The highlighted cases A and B both receive the same number of citations (i.e., two) meaning they have the same in-degree. An eigenvector centrality metric would rank case A better than case B since A is cited by case C, which has a large number of citations.

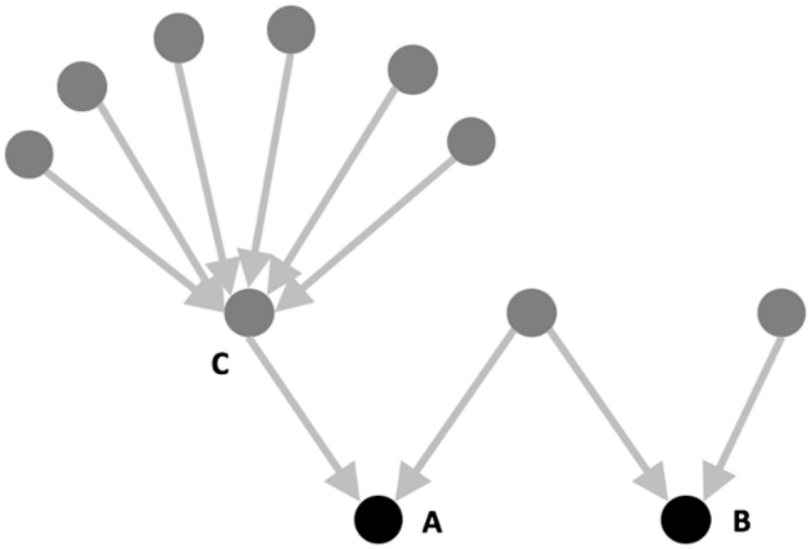



**Figure 5.3.1:** This Figure shows a small citation network. Cases A and B would be ranked equally by in-degree, but case A would be ranked better by eigenvector centrality metrics as described in text above.

Eigenvector centrality measures include: PageRank, Eigenvector centrality[17], and hubs and authorities (Kolaczyk, 2009). PageRank is one of the key mathematical components of Google's search algorithm (Bryan and Leise, 2006). While PageRank works very well for networks of web pages, Section 5.5.4 discusses why it is less appropriate for citation networks.

[16] The degree of a vertex "provides a basic quantification of the extent to which v is connected to other vertices within the graph." (Kolaczyk, 2009)

[17] Confusingly, eigenvector centrality refers to the broader category and a particular member of this category.

The class of metrics that are called positional are based on the idea that important cases are “close" to other cases in the sense that “distance" is measured by the number of citations it would take to go from one case to another (Kolaczyk, 2009). Figure 5.3.2 shows a hypothetical network with seven nodes. The highlighted case A would be ranked highest by positional metrics since it is “closest" to all other nodes on average. Positional metrics include betweenness centrality and closeness centrality.

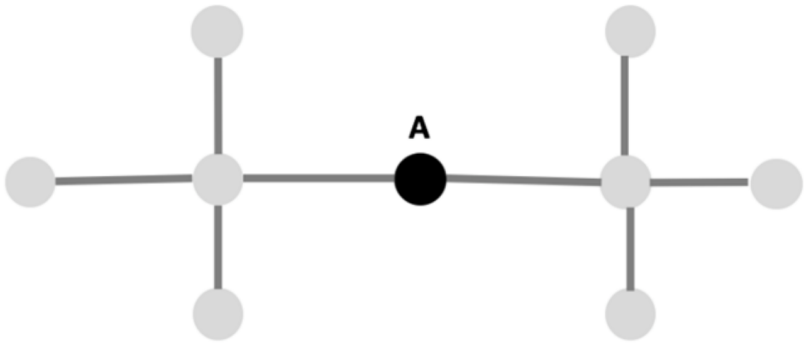



**Figure 5.3.2:** This Figure shows a simple network. The highlighted node A would be ranked highest by most positional vertex centrality metrics since it is “closest" to all other nodes. The network is undirected here for simplicity.

Time plays an important role in the evolution of the legal citation network in that more recent cases are often cited over older cases (Black and Spriggs, 2013). In some circumstances, taking time into account may be desirable. For example, a legal search engine may want to favor newer cases in order to give attorneys quicker access to the most current understanding of the law (Hitt, 2016). Although none of the standard vertex centrality metrics discussed above in this Section incorporate time, it is possible to construct time aware vertex centrality metrics that take the date of each case into account–typically by decreasing the weight of older cases (Walker et al., 2007). Consequently, Section 5.4.1.2 also examines two time-aware vertex centrality metrics: CiteRank and the number of citations in recent years.[18]

### 5.3.2 The methodology

This Comment develops a methodology to compare vertex centrality metrics, which measure some notion of how important a case is in a citation network, based on the evolution of the citation network. The core assumption underlying this methodology is that a better vertex centrality metric will better

[18]There are a number of technical details discussed in Appendix A such as the differences between directed and undirected centrality metrics and the details of the time aware centrality metrics.

predict future citations. In particular, an experiment is run on the citation network, which compares how well each vertex centrality metric can predict future citations.

#### 5.3.2.1 Sort experiment

The experiment attempts to predict which existing cases a new opinion will cite based only on vertex centrality metrics of the citation network at that time. Vertex centrality metrics do not by themselves provide enough information to accurately predict citations; therefore, we slightly modify the problem.[19] Instead of making binary predictions about whether an existing case will be cited by the new case, all existing cases are ranked by vertex centrality metrics. Then, the cases that were actually cited by the new opinion are examined in order to quantify how well these cases were ranked by each vertex centrality metric.

One thousand test cases between 1900 and 2016 were randomly selected to evaluate. For each test case, the citation network just before the test case enters is considered[20] and all vertex centrality metrics of interest (e.g., in-degree, PageRank, etc.) are computed. Each vertex centrality metric gives a ranking of the cases (e.g., in-degree ranks the case with the most citations at this time as the top case). Then, the cases that were actually cited by the test case were observed[21] and the *mean rank score* (Zanin et al., 2009) of these cases was computed for each vertex centrality metric ranking. Mean rank score is related to the average position of the citations in a ranking: **smaller values of mean rank score indicate more predictive ranking**.[22] For each vertex centrality metric the mean rank score was averaged for all one thousand test cases to get an aggregate measure of how predictive of future citations a given vertex centrality metric is. A more detailed discussion of this experiment is provided in Appendix 8.5.

The *sort experiment*[23] described above was run twice, once on the Supreme Court network, and once on the entire Federal Appellate network. For the Supreme Court network, the experiment only looks at citations between Supreme Court cases. In the full Federal Appellate network, citations between Supreme Court cases, between the Supreme Court and the circuit courts, and between the circuit courts are included.

[19] A model that took the topic of the opinion or the judge authoring the opinion into account would make better predictions.

[20] For computational reasons, the network and vertex centrality metrics are computed once each year from 1900 to 2016.

[21] Opinions that cite zero cases are ignored.

[22] There are other ranking metrics such as reciprocal rank. See Appendix 8.5 for a discussion of these choices.

[23] The word sort comes from the fact that the experiment sorts cases by vertex centrality metrics.

#### 5.3.2.2 Motivation and interpretation of the methodology

This methodology has two statistical interpretations. The first interpretation is based on approximating the *link prediction* problem, and the second is based on evaluating the rankings of a *recommender system*.

Link prediction is a problem in network science where one builds a statistical model that attempts to predict future links in network based on current information (e.g., the “people you may know" feature on Facebook).[24] Typically, this problem tries to accurately predict new links based on all available information. The sort experiment approximates this problem in two ways. First, unlike the link prediction model, the sort experiment is based only on vertex centrality metrics and ignores other sources of information such as the topic of the case. Second, link prediction models are typically evaluated by accuracy (e.g., what percent of true future links did the model predict), whereas the sort experiment is evaluated by ranking. This ranking methodology is preferable in this circumstance because one does not expect these predictions to be very accurate[25] and it makes fewer statistical assumptions.[26] Mathematically, one can view a ranking as a relaxation of a probabilistic prediction.

Recommender systems, such as search engines like Google or product recommendations like Netflix or Amazon, rank things to display them in response to a query (e.g., a Google search displays the top ten most relevant results on the first page). The ranking provided by a search engine can be evaluated with some external feedback such as clicks on search results. The sort experiment pretends a new opinion is a “search query" and ranks cases by vertex centrality metrics. The actual citations of the new opinion are then used as feedback to evaluate this ranking.

### 5.3.3 Data

This Comment would not have been possible without the freely available data from CourtListener and the Supreme Court Database (SCDB)(Spaeth et al., 2014).[27] CourtListener describes itself as “a free legal research website containing millions of legal opinions from federal and state courts" that allows

[24] (Fowler et al., 2007) uses a variant of this link prediction methodology where they use vertex centrality metrics to predict whether or not a case will be cited by a given court or in a given year.

[25] Any classification accuracy rate would likely be very low and therefore noisier.

[26] The rankings are non-parametric in the sense that they do not rely on a specific probabilistic model such as logistic regression.

[27] We would also like to thank Mike Lissner, lead developer and co-founder of the Free Law Project, for his willingness to troubleshoot and his clear enthusiasm for quantitative legal research.

"lawyers, journalists, academics, and the public" to "research an important case, stay up to date with new opinions as they are filed, or do deep analysis using our raw data."[28] It contains over three million court opinions from more than 400 jurisdictions and has identified over twenty-five million citations between these opinions.

The SCDB is a freely-accessible database that codifies qualities of Supreme Court cases, including dates of argument and decision, descriptions of litigants, lower court qualities and actions, and issue areas.[29] Widely used by legal scholars (Cross and Spriggs, 2010; Black and Spriggs, 2013; Lupu and Fowler, 2013), it provides a list of Supreme Court cases and a number of pieces of metadata about each case. The SCDB helps mitigate data quality issues from the CourtListener database.

This Comment applies the methodology to two networks: the Supreme Court citation network and the Federal Appellate citation network.[30] Both networks contain cases from 1791 through part of 2016. The Supreme Court network as analyzed contains 27,885 cases and 235,881 citations. For the Supreme Court network, the nodes are Supreme Court opinions and the edges are citations between two Supreme Court cases.

The Federal Appellate network includes cases from the thirteen federal appellate jurisdictions and citations among these opinions. The Federal Appellate network contains 959,985 cases and 6,649,916 citations. The Supreme Court network is a subnetwork of the full Federal Appellate network.

There are several overall limitations to the network data. First, CourtListener only identified whether there is at least one citation between two cases, rather than counting the number of citations between them.[31] Second, CourtListener did not quantify the quality of a citation: whether the citing case follows, distinguishes, or has another relationship to the cited case. Finally, CourtListener grouped all opinions in a case together, so citations in or to a dissent are not distinguished from citations in or to a major-

[28] CourtListener, http://www.courtlistener.com, is an initiative of the Free Law Project, a non-profit "providing free access to primary legal materials, developing legal research tools, and supporting academic research on legal corpora."

[29] Where CourtListener is truly a research tool with a Google-esque search box and the ability to read texts and follow links between them, the Supreme Court Database is a lower level tool, which is generally presented as a spreadsheet or other structured data file.

[30] The numbers of opinions and citations in the following paragraphs reflect the data available from CourtListener as of June 27, 2016, slightly reduced by data processing. The code used to process the data is available at the GitHub site referenced at the beginning of the Appendix.

[31] Most legal citation network analysis shares this limitation (Fowler and Jeon, 2008).

ity opinion or a concurrence.[32] If these data become available in the future, it would be interesting to conduct additional experiments to see if the results change significantly.

## 5.4 Results

This Part first discusses the results of the sort experiment on the Supreme Court network. It first examines in-degree driven metrics, out-degree driven metrics, and then time-aware metrics. It also briefly discusses the results of the sort experiment on the full Federal Appellate network, which are largely similar to the Supreme Court results.

Overall, the results agree with prior work that showed that authorities (an eigenvector-based vertex centrality metric) predicted future citations better than in-degree. The results further show that time-aware metrics predict future citations better than time-agnostic metrics, which also corresponds to prior work. Finally, most surprisingly, the results show that out-degree and related metrics predict future citations more accurately than in-degree and related metrics.

A note about statistical significance for the result: paired t-tests were used to confirm that the differences in mean rank score between vertex centrality metrics are in fact statistically significant. All comparisons that are discussed in the text of this chapter were confirmed to be statistically significant with a significance level alpha = 0.05. This Comment did not control for multiple testing; however, most p-values were very small (order $10^{-10}$ or smaller) so any reasonable multiple testing procedure would likely not have changed the conclusions appreciably.

### 5.4.1 Supreme Court results

This Section considers the result of the sort experiment run on only the Supreme Court network. That is, the vertices of the network are Supreme Court cases, and only citations between Supreme Court cases are considered. The sort experiment in this context compares vertex centrality metrics by how well they predict Supreme Court to Supreme Court citations.

#### 5.4.1.1 Time agnostic metrics

**A. In-degree driven metrics**

[32] For a discussion of how this limitation may have impacted our results, see infra Section 5.5.1.2

Figure 5.4.1 shows the comparison between some of the most commonly used vertex centrality metrics: PageRank, in-degree, authorities, and betweenness centrality. Smaller values of mean rank score indicate better prediction of future citations.

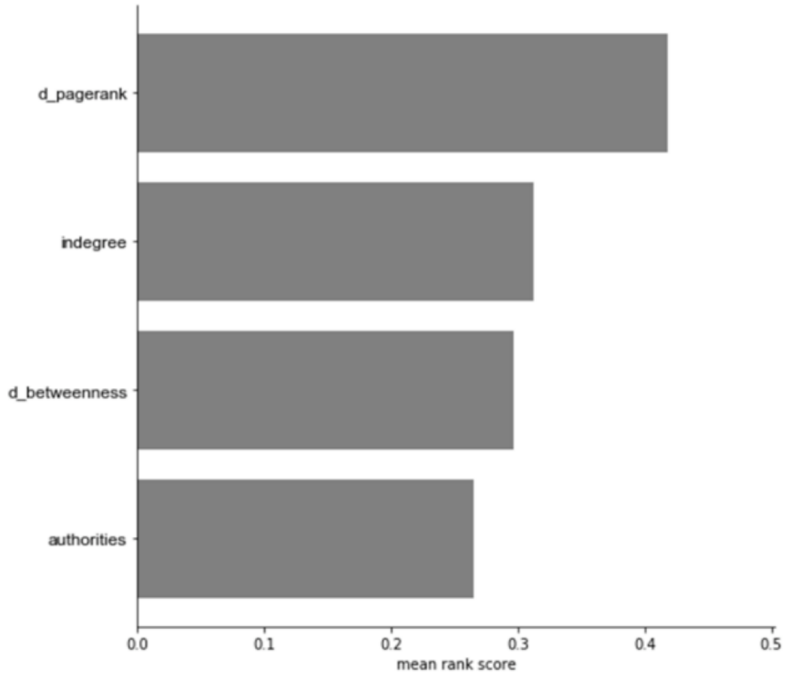



**Figure 5.4.1:** Authorities performs better than the three other in-degree based centrality metrics.

Fowler's *Network Analysis and the Law* compared rankings of Supreme Court cases produced by hubs, authorities, in-degree, and eigenvector based on their ability to predict future citations (Fowler et al., 2007). Pursuant to their methodology, the authors concluded that the authorities score is the vertex centrality metric that best predicts future citations. Consistent with this conclusion, as shown in Figure 5.4.1, authorities beat PageRank, in-degree, and betweenness, which indicates that authorities is more predictive of future citations than these other centrality measures.

**B. Out-degree driven metrics**

Figure 5.4.2 shows the results of the sort experiment for the four in-degree based metrics discussed above and out-degree metrics. Surprisingly, the sort experiment results indicate that out-degree (the

number of cases an opinion cites to) is more predictive of future citations than in-degree (the number of citations an opinion has received).[33]

It seems counterintuitive that the number of cases a judge decided to cite in her opinion would be more predictive of future citations than the number of times other judges have found an opinion worth citing. There are also theoretical mathematical reasons that make this result unexpected, which are discussed in Section 5.5.1.1.

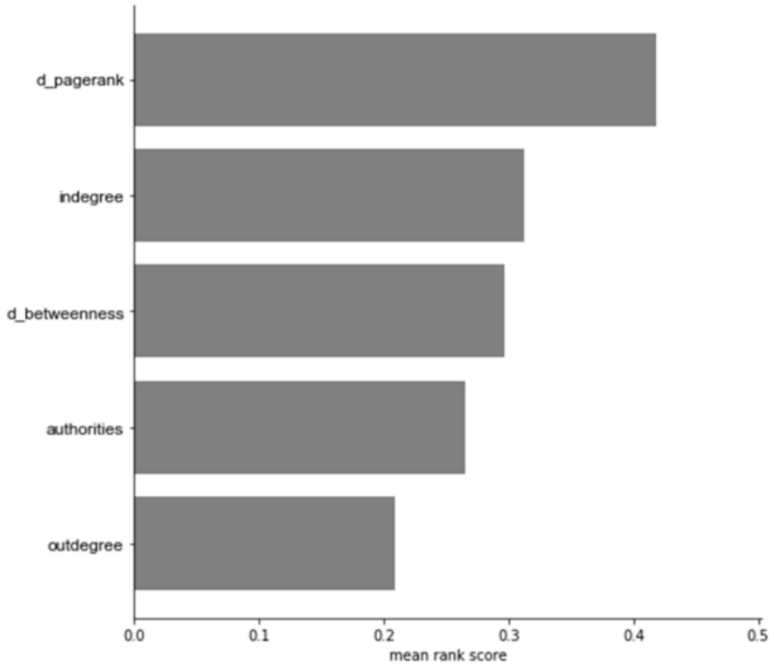



**Figure 5.4.2:** Figure 5.4.2 is the same as Figure 5.4.1 with the addition of out-degree. Out-degree outperforms in-degree and other more sophisticated vertex centrality metrics.

One possible explanation for out-degree's success is that out-degree is a proxy for case length. This explanation can be at least partially investigated with the current data. This Comment measures the length of an opinion by the number of words appearing in the opinion text. A linear regression, shown below in Figure 5.4.3, of number of words versus out-degree resulted in an $R^2$ value of 36%. Thus, out-degree is related to opinion text length. However, opinion text length does not appear to be the sole driver of out-degree.[34]

[33] Additional results relating to vertex centrality metrics, which are driven in part by out-degree, can be found in Appendix 8.5.

[34] Other possible explanations for out-degree's success are discussed in Section 5.5.1.2.

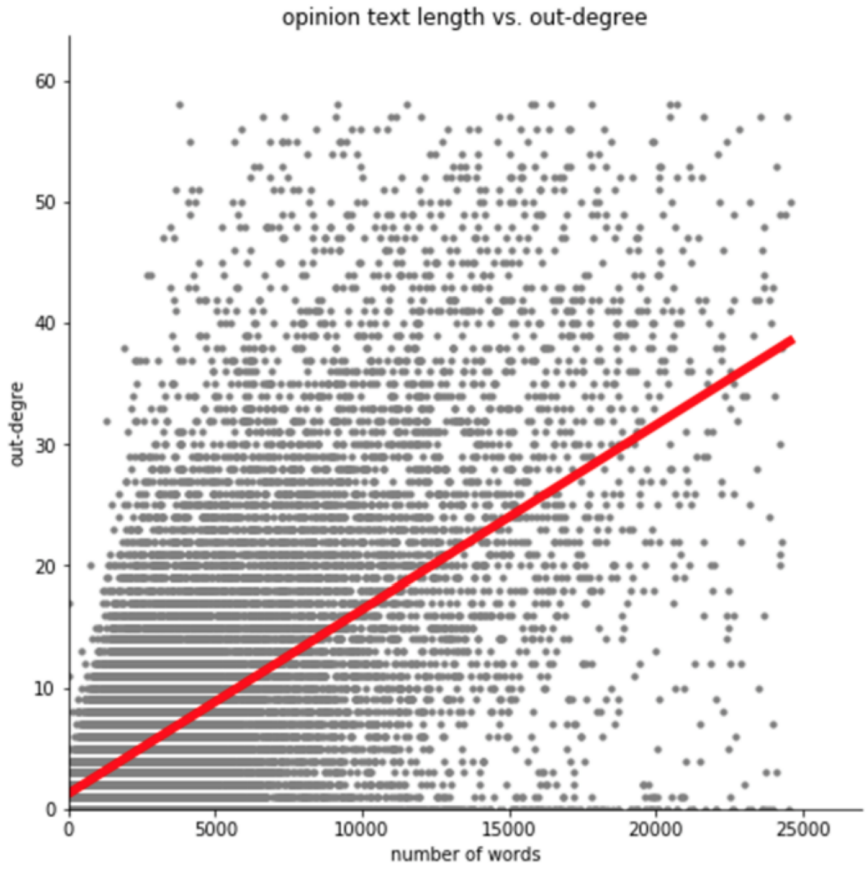



**Figure 5.4.3:** This figure shows a scatter plot of opinion text length and out-degree for all Supreme Court cases. The plot includes the linear model fit of out-degree versus number of words ($R^2 = 0.36$, p-value $< 10^{-3}$). Note that outliers were first removed by removing the top 1% longest cases and top 1% highest out-degree cases. The linear model found a significant relationship at an alpha level of 0.05. The conclusion is that opinion text length and out-degree are related.

Opinion text length was also included in the sort experiment (i.e., cases were ranked by their text lengths). As shown in Figure 5.4.4, opinion text length beats in-degree but does not beat out-degree. To make sure this result is not an artifact of randomness, a pairwise difference t-test found the difference between text length and out-degree to be statistically significant at alpha level = 0.05. Looking at Figure 5.4.4, this difference does appear to be moderate. This result suggests that while opinion text length is important in predicting future citations, out-degree is capturing something beyond just opinion length.

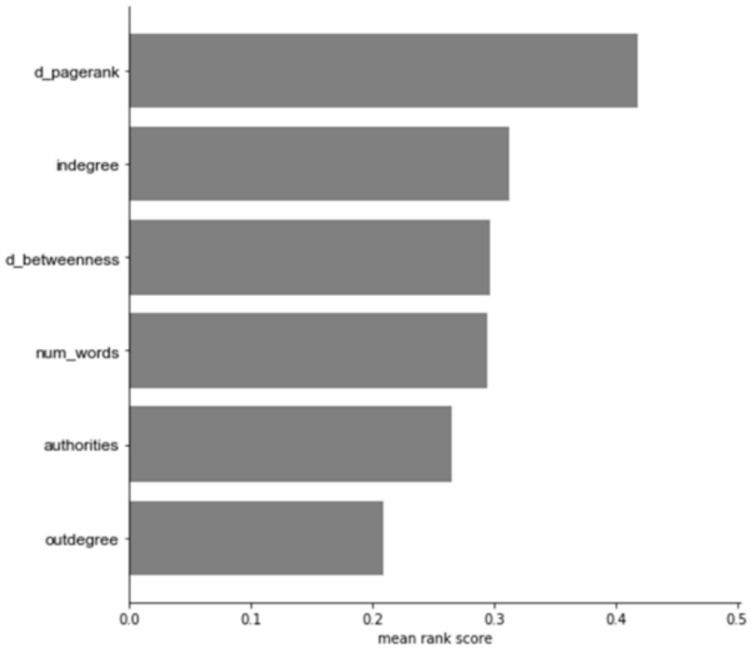



**Figure 5.4.4:** Text length, measured by the number of words contained in the court opinion, does better than in-degree in the sort experiment by a small but statistically significant amount.

#### 5.4.1.2 Time aware metrics

None of the metrics considered so far have taken the age of the case into account. A case that received a large number of citations in the 1920s may not be considered relevant today, but that case will still have a large in-degree value. The data from this experiment and others have shown that cases tend to favor citing recent cases (Black and Spriggs, 2013). For example, Figure 5.4.5 shows a histogram of Supreme Court citation ages

$$\text{citation age} := \text{year of cit}\mathbf{ing}\text{ case} - \text{year of cit}\mathbf{ed}\text{ case}$$

The distribution of citation ages is strongly skewed to the left with a median citation

There is a growing literature on time-aware vertex centrality metrics; however, much of it is beyond the scope of this Comment (Taylor et al., 2017). This Comment evaluates two time-aware vertex centrality metrics: number of citations in the past several years (referred to as RecentCite) and CiteRank (Walker et al., 2007). RecentCite is not a standard name in the networks literature. However, it is a simple way of

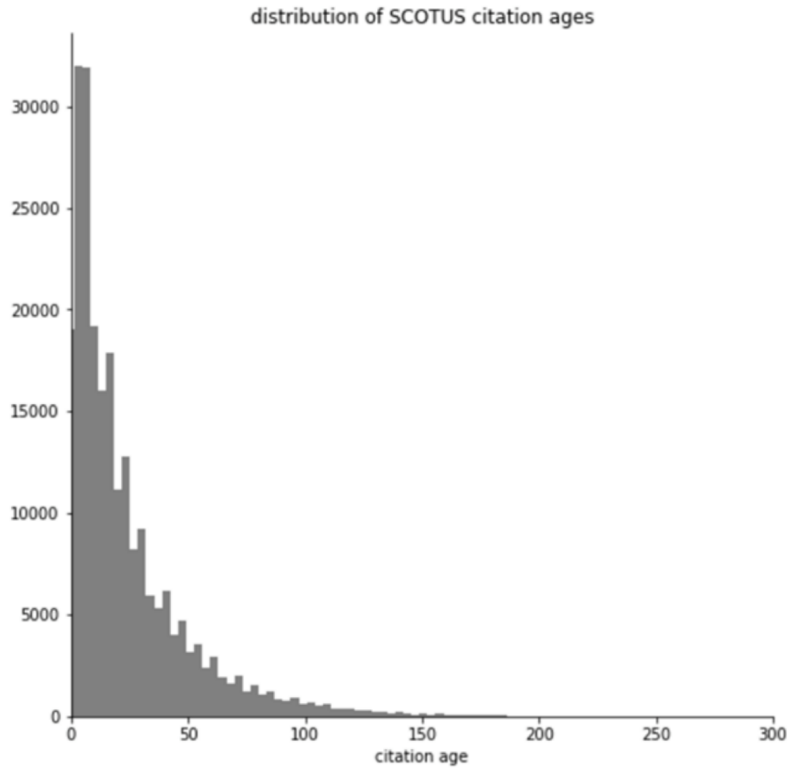



**Figure 5.4.5:** Histogram of citation ages i.e. the difference between the date of the cit*ing* opinion and the cit*ed* opinion. The Supreme Court generally favors citing recent cases over older cases.

measuring how important a case is in recent years. CiteRank, which appears in the networks literature, is a modified version of PageRank that takes the age of a vertex into account and decreases the score for older cases.

Both CiteRank and RecentCite are each really a family of vertex centrality metrics because both have a parameter that controls how heavily older cases are penalized, which can make a significant difference in the case rankings. For example, RecentCite's parameter is simply the cutoff age of whether a citation is counted or excluded. Selecting the value of one of these parameters is beyond the scope of this Comment and is an active statistics research question. Therefore, a range of values is considered for each parameter.

Figure 5.4.6 shows the results of the sort experiment including several time-aware centrality metrics. The age of a case is included as a baseline (i.e., cases are ranked by the date of decision).

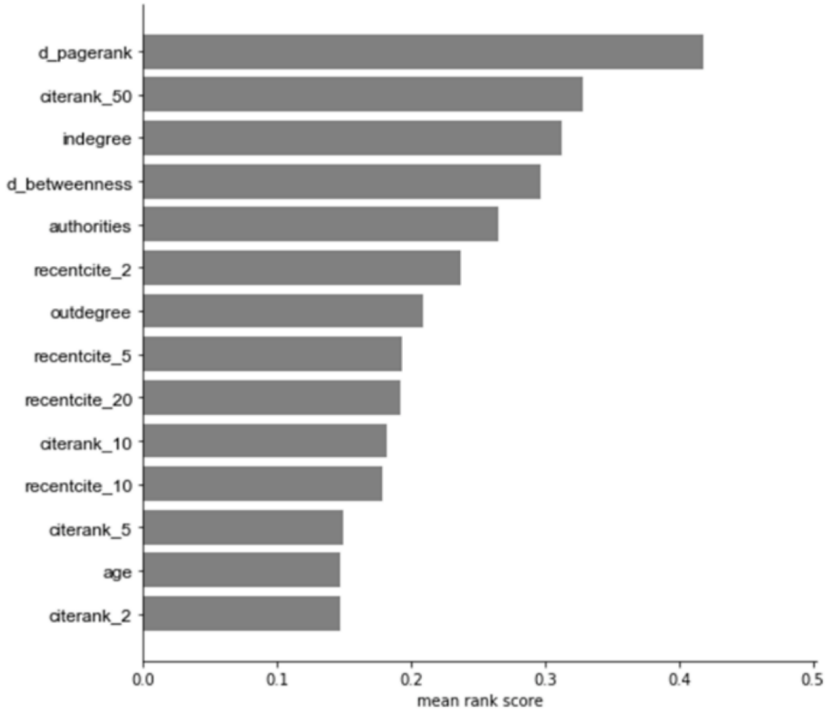



**Figure 5.4.6:** Results of the sort experiment for both time-aware and time-agnostic metrics. Various centrality metrics are on the y-axis and the mean rank score is on the x-axis. Generally, time aware metrics perform better than the time-agnostic metrics.

The first notable feature of Figure 5.4.6 is that most time-aware metrics are better at predicting future citations than the time-agnostic metrics. Based on the citation age distribution in Figure 5.4.5, this may not be surprising: opinions favor citing more recent cases. Therefore, explicitly including a recent time bias will make for better predictions. However, it is surprising that age beats each time-agnostic metric and most time-aware metrics because age ignores all citation information.

Age and RecentCite's performance suggest that many court cases are cited frequently for a period of time soon after they are written and are then cited less frequently. This result is supported by other empirical legal research (Black and Spriggs, 2013; Neale, 2013). Figure 5.4.6 gives some information about the time period in which a case is most likely to be cited. RecentCite ten, which only counts citations from the past ten years, beats both RecentCite two and twenty. The rough conclusion to draw from this is that a case is more likely to be cited when it is more than two and less than twenty years old. This corresponds to the findings of other scholars, one of whom has found that the rate at which a case is cited "depreciate[s] about 81 percent and 85 percent between their first and 20th years of age at the Supreme Court and courts of appeals, respectively."(Black and Spriggs, 2013)

### 5.4.2 Federal Appellate Results

Many of the results discussed above remained broadly the same in the Federal Appellate network sort experiment, with a few exceptions. Figure 5.4.7 shows the results of the sort experiment run on the entire Federal Appellate network.[35] The results for in-degree, out-degree, and authorities are all very close.

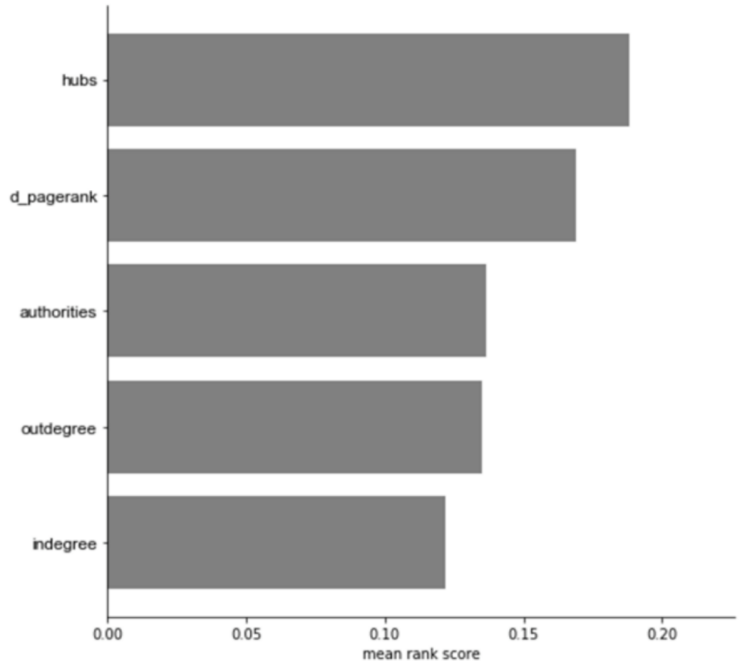



**Figure 5.4.7:** Y-axis is vertex centrality metrics and x-axis is mean rank score. Results of the sort experiment for metrics performed on the Federal Appellate courts: in-degree now does better than both out-degree and authorities. Out-degree beats authorities. Hubs now does worse than in-degree.

In contrast to the results for the Supreme Court, in-degree's predictive power is essentially tied with authorities. Given that in-degree is a simpler metric, this might lead one to prefer in-degree over authorities in the larger network. Furthermore, out-degree is tied with in-degree while hubs score is the worst performing metric. The takeaway is that out-degree still matters, but somewhat less than in the Supreme Court only network.

[35] The results for time-aware metrics were not appreciably different than those for the Supreme Court.

## 5.5 Discussion

Understanding the notion of legal precedent is necessary to understand how the law evolves(Fowler and Jeon, 2008). But the notion of precedent can be difficult to quantify –and, therefore, to study –using empirical methods. The legal citation network, however, provides a natural way of studying precedent because "[e]ach judicial citation contained in an opinion is essentially a latent judgment about the case cited."(Fowler and Jeon, 2008) Building on this theoretical foundation, it is a reasonable assumption that identifying case qualities that are predictive of future citation could yield insights into precedent.

### 5.5.1 Out-degree beats in-degree

The most surprising result of the sort experiment is that out-degree beats in-degree: the number of citations *in* a case (i.e. length of the bibliography) is more predictive of future citations than the number of citations *to* a case. Furthermore, other out-citation centrality metrics (e.g., hubs, reversed PageRank) beat in-citation metrics (e.g., authorities, PageRank).

Because it seems intuitive that cases that have received a lot of citations in the past are likely to receive more citations in the future, in-degree was expected to do well in the sort experiment. In addition to this intuition, there are theoretical reasons discussed below related to why one might expect in-degree to be a good predictor of future citations. Conversely, it seems counterintuitive that the number of cases a judge decided to cite in his opinion would be more predictive of future citations than the number times other judges have found an opinion worth citing, and there is no obvious explanation for why court opinions that cite more cases might be more influential.

Previous scholarship has suggested an association between out-degree and legal relevance.(Cross and Spriggs, 2010) The results of the sort experiment provide additional evidence that the number of citations is strongly associated with future legal relevance. It is not clear why this association exists or which way causation goes.

#### 5.5.1.1 Preferential attachment

One statistical theoretical reason why it is surprising that out-degree is more predictive of future citations than in-degree stems from the concept of preferential attachment. In recent decades, researchers have studied time-evolving networks, such as citation networks, both empirically and theoretically.(Van

Der Hofstad, 2016) Researchers construct mathematical models of how a network evolves that explain features of observed networks (Yule et al., 1925). One popular class of models is called preferential attachment (Barabási and Albert, 1999). The signature feature of a preferential attachment model is that new vertices tend to favor citing cases that already have a lot of citations. Preferential attachment is also referred to as “the rich get richer" phenomenon since cases with a lot of citations will tend to accumulate more citations at a higher rate than cases with fewer citations. Preferential attachment models exhibit topological features[36] that real world networks typically have, such as power law degree distribution (Van Der Hofstad, 2016).[37]

Preferential attachment models are favored by the networks community because they are one of the few types of simple models that exhibit many of the topological features that real world networks, such as the Supreme Court citation network, tend to exhibit.[38] If one assumes the evolutionary dynamics the Supreme Court network obey some kind of preferential attachment model, then one would expect in-degree to be a very strong predictor of future citations. The sort experiment shows that in-degree is not as predictive of future citations as other quantities such as out-degree or case length. This fact suggests that there is possibly some unobserved or latent quantity that is driving the growth of the citation network in a significant way. Understanding what factors are driving the growth of these legal citation networks is an interesting question from both a statistical and legal standpoint.

#### 5.5.1.2 Case qualities possibly driving out-degree

The sort experiment looks at case qualities that are related to the citation network. It is likely that many of these network features are being driven by other case qualities, such as the subject matter of the case, the author of an opinion, or whether a case includes a dissent. Some of these qualities may be related to out-degree.[39] Or, in other words, opinion writers probably do not choose to cite cases on the basis of their out-degree but do chose to cite cases on the basis of qualities that are correlated to out-degree, and it would deepen understanding of precedent to uncover what those qualities are.

[36] A topological feature is a global structural property of a network.

[37] In a power law distribution, a small number of vertices have a large proportion of the total number of edges.

[38] Both the Supreme Court and Federal Appellate networks exhibit a power law distribution of citations.

[39] This is not to say that the qualities discussed in this Comment are the sole possible drivers of out-degree’s performance. For example, the number of legal topics a case addresses could be driving out-degree’s performance. It is also not to say that any one of these qualities is exclusively responsible for out-degree’s performance.

One possible quality that out-degree could reflect is that a case with a higher out-degree is better grounded in existing law. A number of legal scholars have hypothesized that judges prefer citing cases that are better grounded in precedent (Cross and Spriggs, 2010; Lupu and Fowler, 2013). In particular, Fowler and Jeon provide evidence of the relation between how well-grounded a case is and future citations by looking at citations to and from cases during the Warren Court, the period in which Earl Warren was Chief Justice of the Supreme Court (Fowler and Jeon, 2008). This period was marked by novel, highly progressive decisions including *Brown v. Board of Education* (v. Board of Education, 1954), *Miranda v. Arizona* (v. Arizona, 1966), and *Griswold v. Connecticut* (Connecticut, 1965). Relatedly, the Warren Court overruled more precedents than any other Court (Fowler and Jeon, 2008). The fact that the Warren Court broke with existing precedent would mean that their opinions were not grounded in existing law may and potentially reflect that by having lower average out-degree.

As expected, Fowler and Jeon observe that the Warren Court shows a drop in out-degree. “Since the process of creating new law frequently involves breaking with existing precedent, it is no surprise that the Warren Court cited fewer cases in their opinions." (Fowler and Jeon, 2008) However, they also observe that the Warren Court shows a drop in in-degree. In other words, the Warren Court cited fewer Supreme Court opinions and is cited less frequently by future Supreme Court opinions. Fowler and Jeon suggest that the Warren Court’s tendency to break from precedent meant its opinions had “weak legal basis," which is reflected in the drop in out-degree, and which is, in turn, the driving force behind the Warren Court’s lack of citation by subsequent Courts.[40]

The data here show the same patterns. Figure 5.5.1 shows the median in-degree and median out-degree by year for Supreme Court cases. The vertical bars show the timeframe of the Warren Court. During the Warren Court the typical out-degree and the typical in-degree both dip (i.e., future Courts appear to avoid citing Warren Court cases to some extent).

The results of the sort experiment also arguably support the claim at a broader level. If out-degree corresponds generally to how well-grounded an opinion is, then out-degree’s predictive power can be understood as demonstrating a preference in the Supreme Court for citing opinions that are well-grounded

[40] Fowler and Jeon also speculate briefly that the lack of citation to the Warren Court may instead reflect the conservative policy preferences of the Burger and Rehnquist Courts that followed (Fowler and Jeon, 2008).

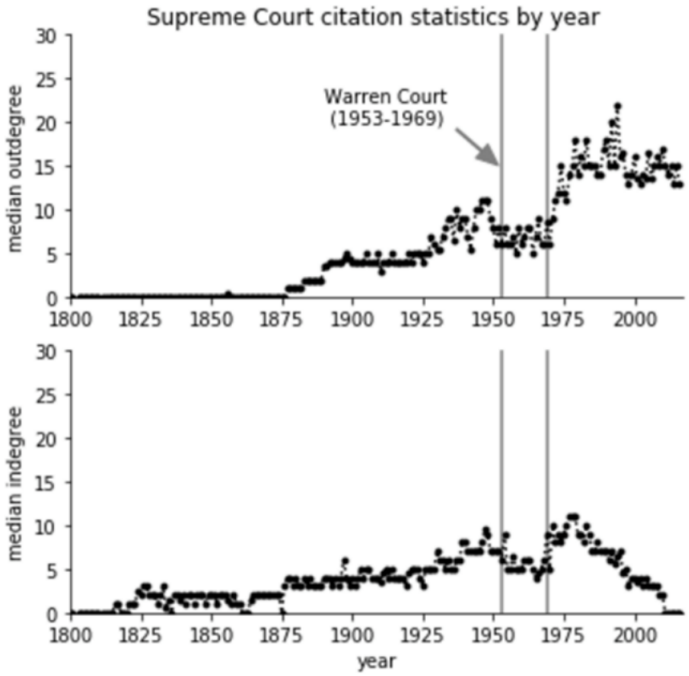



**Figure 5.5.1:** This figure shows the median in-degree and out-degree of Supreme Court cases by year. The Warren Court, which lasted from 1953 to 1969, is visible in the dip in in-degree, out-degree, and case length. Median was selected instead of mean because median is more robust to outliers.

in existing law. The Warren Court correlation between in-degree and out-degree would be part of a larger trend rather than an end in and of itself.

Another quality that may be driving the sort experiment's out-degree results is the presence of dissents and concurrences in a case. Because the data used in the sort experiment groups all opinions in a case together, citations in a dissent or to a dissent are not distinguished from citations in or to a majority opinion or a concurrence. Therefore, a case that contains multiple opinions might have a higher out-degree than a unanimous opinion of equivalent length because opinions coming to a different conclusion, or to the same conclusion but for different reasons, would likely cite to different bodies of cases to support their reasoning.

This observation still leaves unanswered the question of why cases with multiple opinions would be cited more frequently than unanimous opinions. One scholar has demonstrated that the Supreme Court tends to cite “paradoxes"–decisions with a controlling majority as to result–but not as to the grounds of

that result (Hitt, 2016).[41] This tendency reflects the way the Court will incrementally arrive at a rule in a contested area of the law (Hitt, 2016). In contrast, the circuit courts are more likely to cite stable Supreme Court precedent, as opposed to multi-opinion, fractured decisions.

This difference is possibly reflected in the sort experiment. Out-degree was more predictive of future citations than in-degree and more sophisticated metrics in the Supreme Court network. However, in the full Federal Appellate network, out-degree was only as predictive of future citations as in-degree. This could reflect the difference in citation preference between the Supreme Court and the circuit courts. A next step to further explore this possibility would be to re-run the sort experiment on the network of the circuit courts (or to remove citations by the Supreme Court in the full Appellate network), to see if out-degree becomes even less predictive.

Alternatively, it could be that the influence of out-degree is being diluted in the Federal Appellate network by long opinions produced by the circuit courts with very low precedential value, in particular appeals of right in criminal cases with multiple defendants. These cases require very long opinions but are less likely to advance, clarify, or re-shape the law in a substantive way. Table 1 lists the ten cases with the highest out-degree in the Supreme Court network and the full Federal Appellate network. Nine of the ten cases with the highest out-degree in the appellate network are multi-defendant criminal cases, which could indicate that whatever quality is driving the performance of out-degree is being diluted by necessarily long circuit court opinions.

### 5.5.2 Time awareness improves prediction of future citations

The sort experiment also indicated that time-aware centrality metrics are more predictive of future citation than time-agnostic centrality metrics. In other words, incorporating information about how recently a case was decided into a metric improves the ability to predict whether a case will be cited in the future. This result is not surprising but does benefit from further explanation. From a doctrinal perspective, it makes sense that more recent decisions would bear more strongly on current disputes (Landes and Posner, 1976). Given the principle of stare decisis and the analogical processes by which

[41] Hitt describes, *National Mutual Insurance v. Tidewater Transfer*, 337 U.S. 582 (1949), as a quintessential paradox: a majority of justices held that citizens of D.C. could sue in federal court under diversity jurisdiction. But no majority existed on the underlying logic: two justices found in favor of diversity jurisdiction on constitutional grounds, three justices found in favor of jurisdiction on the ability of Congress to grant it (Hitt, 2016).

| Highest out-degree Supreme Court | | Highest out-degree Federal Appellate | | |
|---|---|---|---|---|
| Case name | Year | Case name | Year | Court |
| Miller Brothers Co. v. Maryland | 1954 | United States v. Haldeman | 1977 | D.C. Cir. |
| Commissioner v. Estate of Church | 1949 | United States v. Decoster | 1979 | D.C. Cir. |
| Baker v. Carr | 1962 | United States v. Alvarez | 1987 | 9th Cir. |
| Nebbia v. New York | 1934 | United States v. Byers | 1984 | D.C. Cir. |
| McGautha v. California | 1971 | United States v. Phillips | 1886 | 5th Cir. |
| Crowell v. Benson | 1932 | United States v. Mitchell | 2007 | 9th Cir. |
| Communist Party of United States v. Subversive Activities Control Board | 1961 | United States v. Moore | 2011 | D.C. Cir. |
| The Minnesota Rate Cases | 1890 | United States v. Mitchell | 2007 | 9th Cir. |
| Fay v. Noia | 1963 | United States v. Rigoberto | 1988 | 7th Cir. |
| Oregon v. Mitchell | 1970 | Ruiz v. Estelle | 1982 | 5th Cir. |

**Table 5.5.1:** This table shows the top ten cases by out-degree for the Supreme Court and in the Federal Appellate network. Nine of the top ten cases as ranked by out-degree in the appellate network are multi-defendant criminal cases.

precedent is applied to current disputes, it stands to reason that a more recent decision would capture the nuances of past decisions, providing the most useful “basis for determining later cases involving similar facts or issues." (Landes and Posner, 1976) This principle is often articulated in first-year legal-writing courses, which instruct students to prefer newer cases when selecting authorities to use in an argument (Chew and Pryal, 2016).

Time plays a large role in the evolution of the citation network (e.g., Figure 5.4.5). For legal scholars interested in network analysis, it is worth looking into the growing literature about temporal vertex centrality metrics (Taylor et al., 2017).

### 5.5.3 PageRank and questions of first impression

Given PageRank’s success in ranking web-pages,(Bryan and Leise, 2006) one might expect PageRank to do well in the sort experiment. However, the network topology of the Internet is different from the network topology of a citation network. In a citation network, unlike the Internet, edges can only go in one direction: backwards in time.[42] In other words, while two web pages may link to each other, two cases will only very rarely cite each other.[43] In a network like a citation network, PageRank is known to

[42] Formally, this quality of a citation network makes it a *directed acyclic graph*.

[43] The primary exception to this rule is when the Supreme Court releases two opinions that reference each other on the same day.

be biased in favor of older vertices (Mariani et al., 2016). For an explanation of why this bias occurs, see Appendix 8.5.

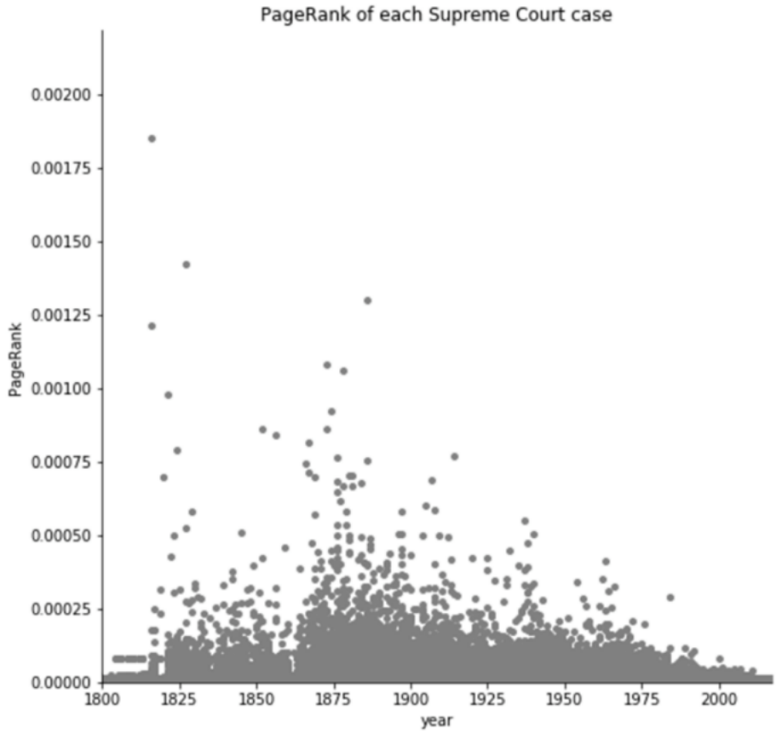



**Figure 5.5.2:** PageRank is biased to favor older cases. This is a plot of each case's PageRank value versus the year that case occurs.

Figure 5.5.2 shows that PageRank favors older cases. The time bias makes PageRank a particularly bad metric for predicting future citations, since, as discussed in Section 5.4.1.2, the Supreme Court prefers citing recent cases. PageRank does, however, seem to pick up on an important quality of case law: questions of first impression. A question of first impression is a legal issue that has not been addressed by the court before. It often involves the first interpretation of a statute or constitutional provision.

Consider the top ten cases ranked by PageRank, listed above. At least four of the top ten presented questions of first impression to the Supreme Court or established fundamental principles of constitutional law. *Gibbons v. Ogden* is the foundational case for interpretation of the Commerce Clause, and *McCulloch v. Maryland* established that the federal government may exercise powers not specifically enumerated in the Constitution. Similarly, *Martin v. Hunter's Lessee* established the Supreme Court's

| Top ten Supreme Court cases by PageRank | |
|---|---|
| Case name | Date decided |
| Gibbons v. Ogden | 1816-03-2 |
| Martin v. Hunter's Lessee | 1816-03-20 |
| McCulloch v Maryland | 1819-03-18 |
| Brown v. Maryland | 1827-03-12 |
| Boyd v. United States | 1886-02-01 |
| Slaughter-House Cases | 1873-04-14 |
| Davidson v. New Orleans | 1878-01-18 |
| Cohens v. Virginia | 1821-03-18 |
| Ex Parte Lange | 1874-01-30 |
| Cooley v. Board of Wardens of Port of Philadelphia | 1852-03-18 |

**Table 5.5.2:** Top ten Supreme Court cases cases by PageRank.

ability to review the decisions of state supreme courts. *The Slaughter-House* Cases contain the Supreme Court's first interpretation of the Fourteenth Amendment, which is at the heart of many of the most famous twentieth and twenty-first century cases regarding individual rights, including *Brown v. Board of Education*, *Roe v. Wade* and *Obergefell v. Hodges*.

### 5.5.4 Deciding which vertex centrality metric to use

As mentioned above, there are a growing number of legal practice areas in which lawyers can benefit from a baseline understanding of the principles underlying legal tools (Lohr, 2017). In some practice areas, such as e-discovery, lawyers may even have to make decisions on technical statistical issues. However, rather than attempting to identify the "best" vertex centrality metric, this Section discusses a range of considerations should be used when deciding which vertex centrality metric to employ in a given circumstance.

As a starting point, an algorithm (such as a vertex centrality metric) will always produce some kind of answer (such as a ranked list of cases). But the fact that the answer has been generated does not indicate that the answer is meaningful or trustworthy. At one level, this Comment interrogates the answers given by a set of algorithms, vertex centrality metrics, and attempts to determine whether the answers given by them are meaningful, and in turn what those answers mean.

But more broadly, the answers produced by the methodology are themselves subject to question. The central assumption of this methodology–that a good vertex centrality metric has the ability to pre-

dict future citations–may not be the most appropriate starting assumption in picking a vertex centrality metric. A legal historical study, for instance, may seek to identify factors driving why cases are consistently cited over time, which this methodology would not help with.[44] Awareness of starting assumptions and qualities of statistical tools help ground the scope of questions that can be asked, and in turn the situations in which a tool might be appropriately used.

To give a more concrete example from the present study, the authorities score is more predictive of future citations than in-degree according to the methodology. However, Occam's razor – the proposition that the simplest solution to a problem is most likely the best one – suggests one should prefer a degree-based metric such as in-degree since they are the simplest metrics unless one of the other metrics performs substantially better. In-degree is less conceptually complicated than authorities; authorities is harder to interpret and requires an understanding of higher math. Additionally, simple algorithms tend to be preferred over complex algorithms, as the more complex an algorithm is, the more things can go wrong because they can be more sensitive to noise and more easily statistically biased (James et al., 2013).

To give another example from this Comment, simply because the methodology shows that out-degree is more predictive of future citations than other measures does not necessarily mean it should be the centrality measure of first choice for a search engine or other predictive tool. Out-degree is simple, but the connection between out-degree and future citations is relatively opaque. For example, if out-degree's predictive performance is relatively unique to the Supreme Court, or if out-degree is a proxy for opinion length or number of topics discussed in an opinion, implementing out-degree as a ranking tool would privilege case qualities that are generally not considered important in the context of legal research.

Outside the context of this Comment, a growing number of e-discovery proceedings employ statistical machine learning processes (Henderson, 2015). Attorneys managing these processes would be better equipped to advise their client and direct employees and contractors with an understanding of the assumptions these statistical processes are built on. And as the number of areas of practice which rely on machine learning and other statistical processes grows, lawyers will be asked to reason about topics like vertex centrality more frequently. [45]

[44] See Section 5.5.3.

[45] From (Lohr, 2017) "[T]he law firm partner of the future will be the leader of a team, 'and more than one of the players will be a machine.'" (quoting Michael Mills, a layer and the Chief Strategy Officer of a legal technology start-up).

## 5.6 Conclusion

This Comment introduces a methodology that evaluates the predictive power of vertex centrality metrics. It hypothesizes latent qualities that could be driving the performance of vertex centrality metrics, such as case length and questions of first impression. A number of potential future research questions, both technical and legal, are discussed in the Appendix.

The sort experiment methodology compares vertex centrality metrics by how well they predict future citations of a case. For a given test case, all previous cases are ranked by a given centrality metric. The methodology then looks at the actual citations of the test cases and where they land in this ranking. The idea is that better centrality metrics will tend to put the cited cases closer to the top of the ranking. The sort experiment is one way of using data to compare vertex centrality metrics. It can be interpreted as evaluating each metric's ability to predict future citations. It can also be interpreted as evaluating a metric's ability to rank cases for a search engine.

The most surprising finding of the sort methodology is that out-degree is more predictive of future citations than in-degree. This result may be evidence for the importance of precedent. It is possible that the number of citations in an opinion (out-degree) is a proxy for how well-grounded in precedent that opinion is. It is also possibly a proxy for whether an opinion contains a dissent or concurrences. Significant additional questions remain in this line of research, including whether the performance of out-degree is unique to the Supreme Court. The methodology could also be improved by use of more nuanced data, such as counting the number of citations between opinions, rather than just the fact of one opinion citing another. Most broadly, citation analysis is a powerful tool that has the potential to illuminate a great deal about the structure and evolution of the law. [46]

[46] We would like to thank Michael Lissner and CourtListener for making this project possible by providing the data; Shankar Bhamidi for encouraging us and helping us to think through this project; David Ardia for early and honest feedback; Deborah and Calum Carmichael for spending a significant amount of time editing the drafts; and Molly Brummett Wudel for being unfailingly supportive.

CHAPTER 6

# Fluctuation bounds for continuous time branching processes and nonparametric change point detection in growing networks

Motivated by applications, both for modeling real world systems as well as in the study of probabilistic systems such as recursive trees, the last few years have seen an explosion in models for dynamically evolving networks. The aim of this chapter is two fold: (a) develop mathematical techniques based on continuous time branching processes (CTBP) to derive quantitative error bounds for functionals of a major class of these models about their large network limits; (b) develop general theory to understand the role of abrupt changes in the evolution dynamics of these models using which one can develop nonparametric change point detection estimators. In the context of the second aim, for fixed final network size $n$ and a change point $\tau(n) < n$, we consider models of growing networks which evolve via new vertices attaching to the pre-existing network according to one attachment function $f$ till the system grows to size $\tau(n)$ when new vertices switch their behavior to a different function $g$ till the system reaches size $n$. With general non-explosivity assumptions on the attachment functions $f, g$, we consider both the standard model where $\tau(n) = \Theta(n)$ as well as the *quick big bang model* when $\tau(n) = n^\gamma$ for some $0 < \gamma < 1$. Proofs rely on a careful analysis of an associated *inhomogeneous* continuous time branching process. Techniques developed in the paper are robust enough to understand the behavior of these models for any sequence of change points $\tau(n) \to \infty$. This chapter derives rates of convergence for functionals such as the degree distribution; the same proof techniques should enable one to analyze more complicated functionals such as the associated fringe distributions.

## 6.1 Introduction

### 6.1.1 Motivation

Driven by the explosion in the amount of data on various real world networks, the last few years have seen the emergence of many new mathematical network models. Motivations behind these models are

diverse including **(a)** extracting unexpected patterns as densely connected regions in the network (e.g. community detection); **(b)** understand properties of dynamics on these real world systems such as the spread of epidemics, the efficacy of random walk search algorithms etc; **(c)** most relevant for this study, understanding mechanistic reasons for the emergence of empirically observed properties of these systems such as heavy tailed degree distribution or the small world property. We refer the interested reader to (Albert and Barabási, 2002; Newman, 2003, 2010; Bollobás, 2001; Durrett, 2007; Van Der Hofstad, 2009) and the references therein for a starting point to the vast literature on network models. A small but increasingly important niche is the setting of *dynamic network models*, networks that evolve over time. In the context of probabilistic combinatorics, in particular in the study of growing random trees, these models have been studied for decades in the vast field of *recursive* trees, see (Mahmoud, 2008; Bergeron et al., 1992; Flajolet and Sedgewick, 2009; Drmota, 2009) and the references therein. To fix ideas, consider one of the standard examples: start with a base graph $\mathscr{G}_0$ (e.g. two vertices connected by an edge) and an attachment function $f : \mathbb{Z}_+ \to [0,\infty)$ where $\mathbb{Z}_+ := \{0,1,2,\ldots\}$. For each fixed time $n \geq 1$, having constructed the network $\mathscr{G}_{n-1}$ at time $n-1$, the network transitions to $\mathscr{G}_n$ as follows: a new vertex enters the system and attaches to a pre-existing vertex $v \in \mathscr{G}_{n-1}$ with probability proportional to $f(\deg(v))$ where $\deg(v)$ is the current degree of this vertex. The case of $f(\cdot) \equiv 1$ corresponds to the famous class of random recursive trees (Smythe and Mahmoud, 1995). The specific case of $f(k) = k \; \forall k \geq 0$ was considered in (Barabási and Albert, 1999) where they showed, via non-rigorous arguments, that the resulting graph has a heavy tailed degree distribution with exponent 3 in the large $n \to \infty$ limit; this was rigorously proved in (Bollobás et al., 2001).

### 6.1.2 Informal description of our aims and results

This chapter has the following two major aims:

(a) In the context of models described above, asymptotics in the large network limit for a host of random tree models as well as corresponding functionals have been derived ranging from the degree distribution to the so-called *fringe* distribution (Aldous, 1991; Holmgren et al., 2017; Bhamidi, 2007) of random trees. One of the major drivers of research has been proving convergence of the empirical distribution of these functionals to limiting (model dependent) constants. Establishing (even sub-optimal) rates of convergence for these models has been non-trivial other than for models related

to urn models e.g. see the seminal work of Janson (Janson, 2004). The aim of this chapter is to develop robust methodology for proving such error bounds for general models. Our results will not be optimal owing to the generality of the model considered in the paper; however using the techniques in this chapter coupled with higher moment assumptions can easily lead to more refined results for specific models. To keep the paper to manageable length, we focus on the degree distribution but see Section 6.4 for our work in progress of using the methodology in this chapter for more general functionals.

(b) Consider general models of network evolution as described in the above paragraph but wherein, beyond some point, new individuals entering the system change their evolution behavior. This is reflected via a change in the the attachment function $f$ to a different attachment function $g$.

(i) We first aim to understand the effect of change points on structural properties of the network model and the interplay between the time scale of the change point and the nature of the attachment functions before and after the change point. Analogous to classical change point detection, we start by considering models which evolve for $n$ steps with a change point at time $\gamma n$ for $0 < \gamma < 1$; we call this the *standard model*. Counter-intuitively, we find that irrespective of the value of $\gamma$, structural properties of the network such as the tail of the degree distribution are determined by model parameters before the change point; motivated by this we consider other time scales of the change point (which we call the *quick big bang* model) to see the effect of the long range dependence phenomenon in the evolution of the process.

(ii) We then develop nonparametric change point detection techniques for the standard model when one has no knowledge of the attachment functions, pre or post change point.

### 6.1.3 Model definition

Fix $J \geq 0$. For each $0 \leq j \leq J$, fix functions $f_j : \mathbb{Z}_+ \to \mathbb{R}_+$, which we will refer to as *attachment* functions. Let us start by describing the model when $J = 0$, and we have one attachment function $f_0$. This setting will be referred to as *nonuniform random recursive trees* (Szymański, 1987) or *attachment model*. We will grow a sequence of random trees $\{\mathcal{T}_j : 2 \leq j \leq n\}$ as follows:

(i) For $n = 2$, $\mathscr{T}_2$ consists of two vertices attached by a single edge. Label these using $\{1,2\}$ and call the vertex $v = 1$ as the "root" of the tree. We will think of the tree as directed with edges being pointed away from the root (from parent to child).

(ii) Fix $n > 2$. Let the vertices in $\mathscr{T}_{n-1}$ be labeled by $[n-1]$. For each vertex $v \in \mathscr{T}_{n-1}$ let out-deg$(v)$ denote the out-degree of $v$. A new vertex labelled by $n$ enters the system. Conditional on $\mathscr{T}_{n-1}$, this new vertex attaches to a currently existing vertex $v \in [n-1]$ with probability proportional to $f_0$(out-deg$(v)$). Call the vertex that $n$ attaches to, the "parent" of $n$ and direct the edge from this parent to $n$ resulting in the tree $\mathscr{T}_n$.

**Model with change point:** Next we define the model with $J \geq 1$ distinct change points. Fix attachment functions $f_0 \neq f_1 \neq f_2 \cdots \neq f_J$. For $n \geq J+2$ fix $J$ distinct times $2 \leq \tau_1 < \tau_2 < \cdots < \tau_J < n \in [n]$. Let $\mathbf{f} = (f_j : 0 \leq j \leq J)$ and $\boldsymbol{\tau} = (\tau_j : 1 \leq j \leq J)$ and write $\boldsymbol{\theta} = (\mathbf{f}, \boldsymbol{\tau})$ for the driving parameters of the process. For notational convenience, let $\tau_0 = 2$ and $\tau_{J+1} = n$. Consider a sequence of random trees $\{\mathscr{T}_i^{\boldsymbol{\theta}} : 2 \leq i \leq n\}$ constructed as follows. For $2 \leq i \leq \tau_1$, the process evolves as in the non-change point model using the attachment function $f_0$. We will call this the *initializer* function. Then or each change point index $1 \leq j \leq J$ and time $i \in (\tau_j, \tau_{j+1}]$ the process evolves according to the function $f_j$ i.e. each new vertex entering the system at time $i \in (\tau_j, \tau_{j+1}]$ attaches to a pre-existing vertex $v \in \mathscr{T}_{i-1}^{\boldsymbol{\theta}}$ with probability proportional to $f_j$(out-deg$(v)$).

### 6.1.4 Organization of the paper

We start by defining fundamental objects required to state our main results in Section 6.2. Our main results are in Section 6.3. In Section 6.4 we discuss the relevance of this work as well as related literature. The remaining sections are devoted to proofs of the main results.

## 6.2 Preliminaries

### 6.2.1 Mathematical notation

We use $\preceq_{\text{st}}$ for stochastic domination between two real valued probability measures. For $J \geq 1$ let $[J] := \{1, 2, \ldots, J\}$. If $Y$ has an exponential distribution with rate $\lambda$, write this as $Y \sim \exp(\lambda)$. Write $\mathbb{Z}$ for the set of integers, $\mathbb{R}$ for the real line and let $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$, $\mathbb{R}_+ := (0, \infty)$. Write $\xrightarrow{\text{a.e.}}, \xrightarrow{P}, \xrightarrow{d}$ for con-

vergence almost everywhere, in probability and in distribution respectively. For a non-negative function $n \mapsto g(n)$, we write $f(n) = O(g(n))$ when $|f(n)|/g(n)$ is uniformly bounded, and $f(n) = o(g(n))$ when $\lim_{n\to\infty} f(n)/g(n) = 0$. Furthermore, write $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$. Finally, we write that a sequence of events $(A_n)_{n\geq 1}$ occurs *with high probability* (whp) when $\mathbb{P}(A_n) \to 1$. For a sequence of increasing rooted trees $\{\mathcal{T}_n : n \geq 1\}$ (random or deterministic), we will assume that edges are directed from parent to child (with the root as the original progenitor). For any $n \geq 1$, note that for all vertices $v \in \mathcal{T}_n$ but the root, the degree of $v$ is the same as the out-degree of $v+1$. For $n \geq 1$ and $k \geq 0$, let $D_n(k)$ be the number of vertices in $\mathcal{T}_n$ with *out-degree k*; thus $D_n(0)$ counts the number of leaves in $\mathcal{T}_n$.

### 6.2.2 Assumptions on attachment functions

Here we setup constructions needed to state the main results. We will need the following assumption on the attachment functions of interest in this chapter. We mainly follow (Jagers, 1975; Jagers and Nerman, 1984a; Nerman, 1981; Rudas et al., 2007).

**Assumption 6.2.1.** (i) **Positivity:** *Every attachment function $f$ is assumed to be strictly positive that is $f(k) > 0$ for all $k$.*

(ii) *Every attachment function $f$ can grow at most linearly i.e. $\exists C < \infty$ such that $\limsup_{k\to\infty} f(k)/k \leq C$. This is equivalent to there existing a constant $C$ such that $f(k) \leq C(k+1)$ for all $k \geq 0$.*

(iii) *Consider the following function $\hat{\rho} : (0,\infty) \to (0,\infty]$ defined via,*

$$\hat{\rho}(\lambda) := \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{f(i)}{\lambda + f(i)}. \tag{6.2.1}$$

*Define $\underline{\lambda} := \inf\{\lambda > 0 : \hat{\rho}(\lambda) < \infty\}$. We assume,*

$$\lim_{\lambda \downarrow \underline{\lambda}} \hat{\rho}(\lambda) > 1. \tag{6.2.2}$$

Using (iii) of the above Assumption, let $\lambda^* := \lambda^*(f)$ be the unique $\lambda$ such that

$$\hat{\rho}(\lambda^*) = 1. \tag{6.2.3}$$

This object is often referred to as the Malthusian rate of growth parameter.

### 6.2.3 Branching processes

Fix an attachment function $f$ as above. We can construct a point process $\xi_f$ on $\mathbb{R}_+$ as follows: Let $\{E_i : i \geq 0\}$ be a sequence of independent exponential random variables with $E_i \sim \exp(f(i))$. Now define $L_i := \sum_{j=0}^{i-1} E_i$ for $i \geq 1$. The point process $\xi_f$ is defined via,

$$\xi_f := (L_1, L_2, \ldots). \tag{6.2.4}$$

Abusing notation, we write for $t \geq 0$,

$$\xi_f[0,t] := \#\{i : L_i \leq t\}, \qquad \mu_f[0,t] := \mathbb{E}(\xi_f[0,t]). \tag{6.2.5}$$

Here we view $\mu_f$ as a measure on $(\mathbb{R}_+, \mathscr{B}(\mathbb{R}_+))$. We will need a variant of the above objects: for fixed $k \geq 0$, let $\xi_f^{(k)}$ denote the point process where the first inter-arrival time is $E_k$ namely define the sequence,

$$L_i^{(k)} = E_k + E_{k+1} + \cdots E_{k+i-1}, \qquad i \geq 1.$$

Then define,

$$\xi_f^{(k)} := (L_1^{(k)}, L_2^{(k)}, \ldots), \qquad \mu_f^{(k)}[0,t] := \mathbb{E}(\xi_f^{(k)}[0,t]). \tag{6.2.6}$$

As above, $\xi_f^{(k)}[0,t] := \#\left\{i : L_i^{(k)} \leq t\right\}$. We abbreviate $\xi_f[0,t]$ as $\xi_f(t)$ and similarly $\mu_f(t), \xi_f^{(k)}(t), \mu_f^{(k)}(t)$.

**Definition 6.2.2** (Continuous time Branching process (CTBP))**.** *Fix attachment function $f$ satisfying Assumption 6.2.1(ii). A continuous time branching process driven by $f$, written as $\{\mathrm{BP}_f(t) : t \geq 0\}$, is defined to be a branching process started with one individual at time $t = 0$ and such that every individual born into the system has an offspring distribution that is an independent copy of the point process $\xi_f$ defined in* (6.2.4).

We refer the interested reader to (Jagers, 1975; Athreya and Ney, 1972) for general theory regarding continuous time branching processes. We will also use $\mathrm{BP}_f(t)$ to denote the collection of all individuals at time $t \geq 0$. For $x \in \{\mathrm{BP}_f(t) : t \geq 0\}$, denote by $\sigma_x$ the birth time of $x$. Let $Z_f(t)$ denote the size (number of individuals born) by time $t$. Note in our construction, by our assumption on the attachment function,

individuals continue to reproduce forever. Write $m_f(\cdot)$ for the corresponding expectation i.e.,

$$m_f(t) := \mathbb{E}(Z_f(t)), \qquad t \geq 0, \tag{6.2.7}$$

Under Assumption 6.2.1(ii), it can be shown (Jagers, 1975, Chapter 3) that for all $t > 0$, $m_f(t) < \infty$, is strictly increasing with $m_f(t) \uparrow \infty$ as $t \uparrow \infty$. In the sequel, to simplify notation we will suppress dependence on $f$ on the various objects defined above and write $\mathrm{BP}(\cdot), m(\cdot)$ etc. The connection between CTBP and the discrete random tree models in the previous section is given by the following result which is easy to check using properties of exponential distribution (and is the starting point of the Athreya-Karlin embedding (Athreya and Karlin, 1968)).

**Lemma 6.2.3.** *Fix attachment function $f$ consider the sequence of random trees $\{\mathcal{T}_m : 2 \leq m \leq n\}$ constructed using attachment function $f$. Consider the continuous time construction in Definition 6.2.2 and define for $m \geq 1$ the stopping times $T_m := \inf\{t \geq 0 : |\mathrm{BP}_f(t)| = m\}$. Then viewed as a sequence of growing random labelled rooted trees we have, $\{\mathrm{BP}_n(T_m) : 2 \leq m \leq n\} \stackrel{d}{=} \{\mathcal{T}_m : 2 \leq m \leq n\}$.*

### 6.3 Main Results

#### 6.3.1 Convergence rates for model without change point

Consider a continuous time branching process with attachment function $f$ and Malthusian rate $\lambda^*$. For each $k \geq 0$, $t \geq 0$, denote by $D(k,t)$ the number of vertices in $\mathrm{BP}_f(t)$ of degree $k$ and abbreviate $Z_f(t)$ to $Z(t)$. Let $\lambda^* = \lambda^*(f)$ be as in (6.2.3). Define the probability mass function $\mathbf{p}(f) := \{p_k : k \geq 0\}$ via,

$$p_k = p_k(f) := \frac{\lambda^*}{\lambda^* + f(k)} \prod_{j=0}^{k-1} \frac{f(i)}{\lambda^* + f(i)}, \qquad k \geq 0. \tag{6.3.1}$$

Here for $k = 0$, the $\prod_{j=0}^{k-1}$ is by convention taken to be 1. Verification that the above is an honest probability mass function can be found in (Rudas et al., 2007, Theorem 2). Following the seminal work of (Jagers, 1975; Jagers and Nerman, 1984a; Nerman, 1981; Rudas et al., 2007), it follows that for each $k \geq 0$ that

$$\frac{D(k,t)}{Z(t)} \xrightarrow{\mathrm{P}} p_k \quad \text{as} \quad t \to \infty.$$

However, to obtain consistent change point estimators, we need to strengthen the above convergence to a sup-norm convergence on a time interval whose size goes to infinity with growing $t$ and also, a quantitative rate for this convergence. Such results have been obtained for very specific attachment functions via functional central limit theorems (e.g. see (Janson, 2004) for models whose degree evolution can be reduced to the evolution of urn processes satisfying regularity conditions and (Resnick and Samorodnitsky, 2015) for the linear preferential attachment model), but do not extend to the general setting. We make the following assumptions throughout this section.

**Assumption 6.3.1.** *There exists* $C^* \geq 0$ *such that* $\lim_{k\to\infty} f(k)/k = C^*$.

**Assumption 6.3.2.** $\mathrm{Var}\left(\int_0^\infty e^{-\lambda^* t}\xi_f(dt)\right) < \infty$.

**Remark 1.** *Assumption 6.3.2 is implied by* $\sum_{k=0}^\infty k^2 p_k(f) < \infty$ *since*

$$\mathbb{E}\left(\int_0^\infty e^{-\lambda^* t}\xi_f(dt)\right)^2 = \mathbb{E}\left(\int_0^\infty \lambda^* e^{-\lambda^* t}\xi_f(t)dt\right)^2 \leq \mathbb{E}\left(\int_0^\infty \lambda^* e^{-\lambda^* t}\xi_f^2(t)dt\right) = \sum_{k=1}^\infty k^2 p_k(f_0) < \infty.$$

Fix a sequence of growing trees $\{\mathcal{T}_m : m \geq 2\}$ and recall that for any $N \geq 2$ and $k \geq 0$, $D_N(k)$ denotes the number of vertices with out-degree $k$. The main theorem of this section is

**Theorem 6.3.3.** *Consider a continuous time branching process with attachment function* $f$ *that satisfies Assumptions 6.2.1, 6.3.1 and 6.3.2. Let* $(p_1, p_2, \ldots)$ *denote the limiting degree distribution. There exist* $\omega^*, \epsilon^{**} \in (0,1)$, *such that for any* $\epsilon \leq \epsilon^{**}$,

$$n^{\omega^*}\sum_{k=0}^\infty 2^{-k}\left(\sup_{t\in[0,2\epsilon\log n/\lambda^*]}\left|\frac{D\left(k,\frac{1-\epsilon}{\lambda^*}\log n + t\right)}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n + t\right)} - p_k\right|\right) \xrightarrow{P} 0.$$

*Thus for a sequence of nonuniform recursive trees* $\{\mathcal{T}_m : m \geq 2\}$ *grown using attachment function* $f$,

$$n^{\omega^*}\sum_{k=0}^\infty 2^{-k}\left(\sup_{n^{1-\varepsilon}\leq N\leq n^{1+\varepsilon}}\left|\frac{D_N(k)}{N} - p_k\right|\right) \xrightarrow{P} 0.$$

**Remark 2.** *In the notation of Jagers and Nerman (Jagers and Nerman, 1984b; Nerman, 1981), the result above is stated for the "characteristic" corresponding to degree (see the discussion below). We believe our proof techniques are robust enough to generalize to more complex functionals such as the fringe distribution (Aldous, 1991; Holmgren et al., 2017). We will pursue this in a separate paper. However below we describe one of the key estimates derived in this chapter of more general relevance.*

**Remark 3.** *For special cases such as the uniform or linear preferential attachment, stronger results are obtainable via Janson's "superball" argument (Janson, 2004) as well as application of the Azuma-Hoeffding inequality (Bollobás et al., 2001; Van Der Hofstad, 2009). However these do not appear to work for the general model considered in this chapter.*

Recall from (Nerman, 1981) that a characteristic $\phi$ is a non-negative random process $\{\phi(t) : t \in \mathbb{R}\}$, assigning some kind of score to the typical individual at age t. We assume $\phi(t) = 0$ for every $t < 0$. For this article, we will be interested in the following class of characteristics:

$$\mathscr{C} := \{\phi \text{ with càdlàg paths } : \exists\, b_\phi > 0 \text{ such that } \phi(t) \le b_\phi(\xi_f(t) + 1) \text{ for all } t \ge 0\}. \tag{6.3.2}$$

For any characteristic $\phi$, define $Z_f^\phi(t) := \sum_{x \in \mathrm{BP}_f(t)} \phi(t - \sigma_x)$. This can be thought of as the sum of $\phi$-scores of all individuals in $\mathrm{BP}_f(t)$. Write $m_f^\phi(t) = \mathbb{E}(Z_f^\phi(t))$ $M_f^\phi(t) = \mathbb{E}\left(e^{-\lambda^* t} Z_f^\phi(t)\right)$. For fixed $k \ge 0$ and for the specific characteristic $\phi(t) = \mathbb{1}\{\xi(t) = k\}$, write $m_f^{(k)}(\cdot) := m_f^\phi(\cdot)$.

It is easy to check that for a general (integrable) characteristic $\phi$, $M_f^\phi(t)$ satisfies the renewal equation

$$M_f^\phi(t) = e^{-\lambda^* t}\mathbb{E}(\phi(t)) + \int_0^t M_f^\phi(t-s)e^{-\lambda^* s}\mu_f(ds). \tag{6.3.3}$$

Write $M_f^\phi(\infty) = \lim_{t\to\infty} M_f^\phi(t)$ when the limit exists. Following (Nerman, 1981), we write $x = (x', i)$ to denote that $x$ is the $i$-th child of $x'$ and define for any $t \ge 0$,

$$\mathscr{I}(t) = \{x = (x', i) : \sigma_{x'} \le t \text{ and } t < \sigma_x < \infty\}.$$

Write $W_t = \sum_{x \in \mathscr{I}(t)} e^{-\lambda^* \sigma_x}$. By Corollary 2.5 of (Nerman, 1981), $W_t$ converges almost surely to a finite random variable $W_\infty$ as $t \to \infty$. By Theorem 3.1 of (Nerman, 1981), $e^{-\lambda^* t} Z_f^\phi(t) \xrightarrow{P} W_\infty M_f^\phi(\infty)$ for any $\phi \in \mathscr{C}$. An important technical contribution of this chapter is the following result.

**Theorem 6.3.4.** *Consider a continuous time branching process with attachment function $f$ that satisfies Assumptions 6.2.1, 6.3.1 and 6.3.2. There exist positive constants $C_1, C_2$ such that for any $b_\phi > 1$ and any characteristic $\phi \in \mathscr{C}$ satisfying $|\phi(t)| \le b_\phi(\xi_f(t) + 1)$ for all $t \ge 0$,*

$$\mathbb{E}\left|e^{-\lambda^* t} Z_f^\phi(t) - W_\infty M_f^\phi(\infty)\right| \le C_1 b_\phi e^{-C_2 t}.$$

**Remark 4.** *The constants $\omega^*$ in Theorem 6.3.3 and $C_1, C_2$ in Theorem 6.3.4 are explicitly computable from our proof techniques. However, they depend on the Malthusian rate and thus we have not tried to derive an explicit form of these objects.*

### 6.3.2 Change point detection: Sup-norm convergence of degree distribution for the standard model

Fix $J \geq 1$. We start by studying the model under the following assumption which we refer to as the "standard" model owing to the analogous assumptions for change point methodology in time series:

**Assumption 6.3.5.** *Fix $J \geq 1$ and assume there exist $0 < \gamma_1 < \cdots < \gamma_J < 1$ such that for all $1 \leq j \leq J$, the $j^{th}$ change point is $\tau_j = \lfloor n\gamma_j \rfloor$.*

To simplify notation we will drop $\lfloor\,\rfloor$. Recall the sequence of random trees $\{\mathscr{T}_m^{\boldsymbol{\theta}} : 2 \leq m \leq n\}$. For any $0 < t \leq 1$ and $k \geq 0$, write $D_n(k, nt)$ for the number of vertices with out-degree $k$. We will sometimes abuse notation and write $D_n(k, \mathscr{T}_{nt}) := D_n(k, nt)$ to explicitly specify the dependence of this object on the underlying tree. In this section we mainly consider the case where there is exactly one change point at time $n\gamma_1$ for fixed $0 < \gamma_1 < 1$. In Section 6.3.3 we describe the general result for multiple change points. The notation is cumbersome so this general case can be skipped over on an initial reading. We also give the proof for the single change point case; the general case follows via straight-forward extensions. Fix initializer attachment function $f_0$ and let $\lambda_0^* = \lambda^*(f_0)$ be as in (6.2.3). Define the probability mass function $\{p_k^0 : k \geq 0\}$ via (6.3.1) with $(\lambda_0^*, f_0)$ in place of $(\lambda^*, f)$. As before let the attachment function after change point be $f_1$. Recall from (6.2.6), for fixed $k \geq 0$, the function $\mu_{f_1}^{(k)}[0, \cdot]$ and the function $m_{f_1}(\cdot)$ from (6.2.7). Also recall that, for fixed $k \geq 0$,

$$m_{f_1}^{(k)}(t) = \mathbb{E}\left( \sum_{x \in \mathrm{BP}_{f_1}(t)} \mathbb{1}\left\{ \xi_{f_1}(t - \sigma_x) = k \right\} \right).$$

It can be checked (using the continuity estimates in obtained in Lemmas 6.6.2 and 6.6.9 that for any $k \geq 0$, $t \geq 0$, $m_{f_1}^{(k)}(t) = \int_0^t \mathbb{P}\left(\xi_{f_1}(u) = k\right) m_{f_1}(t - du)$. For $\ell, k \geq 0$, define

$$\lambda_\ell(t) = 1 + \int_0^t m_{f_1}(t-s)\mu_{f_1}^{(\ell)}(ds), \quad \lambda_\ell^{(k)}(t) = \mathbb{P}\left(\xi_{f_1}^{(l)}(t) = k - \ell\right) + \int_0^t m_{f_1}^{(k)}(t-s)\mu_{f_1}^{(\ell)}(ds). \tag{6.3.4}$$

Let $\mathscr{P}$ denote the collection of all probability measures on $\mathbb{N} \cup \{0\}$. For each $a > 0$, consider the functional $\Phi_a : \mathscr{P} \to \mathscr{P}$ given by

$$\Phi_a(\mathbf{p}) = \left( \frac{\sum_{\ell=0}^{\infty} p_\ell \lambda_\ell^{(k)}(a)}{\sum_{\ell=0}^{\infty} p_\ell \lambda_\ell(a)} \right)_{k \geq 0} \tag{6.3.5}$$

where $\mathbf{p} = (p_0, p_1, \ldots) \in \mathscr{P}$. Write $(\Phi_a(\mathbf{p}))_k$ for the $k$-th co-ordinate of the above map. Let $\mathbf{p}^i = \mathbf{p}(f_i) := (p_0^i, p_1^i, \ldots)$ for $i = 0, 1$ denote the degree distribute or a random recursive tree grown with attachment function $f_i$ (i.e. without any change point). Corollary 6.7.2 shows that for each $t > \gamma$, there is a unique $0 < a_t < \infty$ such that

$$\sum_{k=0}^{\infty} p_k^0 \left[ \int_0^{a_t} m_{f_1}(a_t - s) \mu_{f_1}^{(k)}(ds) \right] = \frac{t - \gamma}{\gamma}. \tag{6.3.6}$$

Define $a_t = 0$ for $t \leq \gamma$. Now, we are ready to state our main theorem on sup-norm convergence of degree distributions post-change point.

**Theorem 6.3.6.** *Suppose $f_0, f_1$ satisfy Assumption 6.2.1. For any $k \geq 0$ and $s \in [\gamma, 1]$*

$$\sup_{t \in [\gamma, s]} \left| \frac{D_n(k, nt)}{nt} - (\Phi_{a_t}(\mathbf{p}^{\mathbf{0}}))_k \right| \xrightarrow{\mathrm{P}} 0.$$

There is a probabilistic way to view the limit which we now describe at the end of the construction of the process namely $t = 1$. Write $\alpha$ for $a_1$. Construct an integer valued random variable $D_{\boldsymbol{\theta}}$ using the following auxiliary random variables:

**Construction 6.3.7** ($X_{\mathsf{BC}}$)**.** *Generate $D \sim \{p_k^0 : k \geq 1\}$. Conditional on $D = k$, generate point process $\xi_{f_1}^{(k)}$ and let $\mathfrak{C} = \xi_{f_1}^{(k)}[0, \alpha]$, with $\alpha$ as in* (6.3.6)*. Now set $X_{\mathsf{BC}} = D + \mathfrak{C}$.*

**Construction 6.3.8** ($X_{\mathsf{AC}}$, Age)**.** (a) *Generate $D \sim \{p_k^0 : k \geq 1\}$.*

(b) *Conditional on $D = k$, generate random variable* Age *supported on the interval $[0, \alpha]$ with distribution*

$$\mathbb{P}(\mathsf{Age} > u) := \frac{\int_0^{\alpha - u} m_{f_1}(\alpha - u - s) d\mu_{f_1}^{(k)}(ds)}{\int_0^{\alpha} m_{f_1}(\alpha - s) \mu_{f_1}^{(k)}(ds)}, \qquad 0 \leq u \leq \alpha. \tag{6.3.7}$$

(c) *Conditional on $D$ and* Age*, let $X_{\mathsf{AC}} = \xi_{f_1}[0, \mathsf{Age}]$, where as in* (6.2.5)*, $\xi_{f_1}$ is the point process constructed using attachment function $f_1$.*

Now let $\boldsymbol{\theta} = ((f_0, f_1), \gamma)$. Let $D_{\boldsymbol{\theta}}$ be the integer valued random variable defined as follows: with probability $\gamma$, $D_{\boldsymbol{\theta}} = X_{\mathsf{BC}}$ and with probability $1-\gamma$, $D_{\boldsymbol{\theta}} = X_{\mathsf{AC}}$. The following is a restatement of the convergence result implied by Theorem 6.3.6 for time $t = 1$.

**Theorem 6.3.9** (Standard model, $J = 1$)**.** *As in Section 6.2, fix $k \geq 0$ and let $D_n(k)$ denote the number of vertices with out-degree $k$ in the tree $\mathcal{T}_n^{\boldsymbol{\theta}}$. Under Assumption 6.2.1 on the attachment functions $f_0, f_1$ and Assumption 6.3.5 on the change point $\gamma$, we have that*

$$\frac{D_n(k)}{n} \xrightarrow{P} \mathbb{P}(D_{\boldsymbol{\theta}} = k).$$

Write $\mathbf{p}(\boldsymbol{\theta})$ for the pmf of $D_{\boldsymbol{\theta}}$. The next result, albeit intuitively reasonable is non-trivial to prove in the generality of the models considered in the paper.

**Corollary 6.3.10.** *Assume that $\mathbf{p}^0 \neq \mathbf{p}^1$. Then for any $0 < \gamma < 1$ one has $\mathbf{p}^0 \neq \mathbf{p}(\boldsymbol{\theta})$. Thus the change point always changes the degree.*

The next result describes the tail behavior of the ensuing random variable.

**Corollary 6.3.11** (Initializer always wins)**.** *The initializer function $f_0$ determines the tail behavior of $D_{\boldsymbol{\theta}}$ in the sense that*

(i) *If in the model without change point using $f_0$, the degree distribution has an exponential tail then so does the model with change point irrespective of $\gamma > 0$ and $f_1(\cdot)$.*

(ii) *If in the model without change point using $f_0$, the degree distribution has a power law tail with exponent $\kappa > 0$ then so does model with change point irrespective of $\gamma > 0$ and $f_1(\cdot)$.*

**Corollary 6.3.12** (Maximum degree)**.** *Suppose the initializer function is linear with $f_0(i) = i + 1 + \alpha$ for $i \geq 0$. For fixed $k \geq 1$, let $M_n(k)$ denote the size of the $k$-th maximal degree. Then as long as the function $f_1$ satisfies Assumption 6.2.1, $M_n(k)/n^{1/(\alpha+2)}$ is a tight collection of random variables bounded away from zero as $n \to \infty$.*

**Remark 5.** *Without change point, it is known (Móri, 2007) that for each fixed $k$, $M_n(k)/n^{1/(\alpha+2)} \xrightarrow{d} X_k(\alpha)$ for a non-degenerate distribution. Thus the above result shows that irrespective of the second attachment function $f_0$, the maximal degree asymptotics for linear preferential attachment remain unaffected. The*

*proof of the above result follows via analogous arguments as (Bhamidi et al., 2015, Theorem 2.2) and thus we will not prove it in this chapter.*

### 6.3.3 Multiple change points

Fix $J \geq 1$, $\boldsymbol{\gamma} := (\gamma_1, \gamma_2, \ldots, \gamma_J)$ with $0 < \gamma_1 < \gamma_2 < \cdots < \gamma_J < 1$ and let $\gamma_0 = 0, \gamma_{J+1} = 1$. Further fix attachment functions $f_0, f_1, \ldots f_J$ satisfying Assumption 6.2.1 and let $\boldsymbol{f} := (f_0, f_1, \ldots, f_J)$. We start with the following recursive construction of a sequence of probability mass functions $\{\mathbf{p}^j : 0 \leq j \leq J\}$ and positive constants $\boldsymbol{\alpha} := \{\alpha_j : 1 \leq j \leq J\}$.

(a) **Initialization:** For $j = 0$. let $\mathbf{p}^0 := \{p_k^0 : k \geq 0\}$ as in (6.3.1).

(b) **Pre-epoch distribution:** For $1 \leq j \leq J+1$, define the random variable $X_{PE}^{j-1} \sim \mathbf{p}^{j-1}$.

(c) $\alpha$ **recursion:** For $1 \leq j \leq J$, define $\alpha_j > 0$ as the unique root of the equation:

$$\sum_{k=0}^{\infty} p_k^{(j-1)} \left[ \int_0^{\alpha_j} m_{f_j}(\alpha - s) d\mu_{f_j}^{(k)}(s) \right] := \frac{\gamma_{j+1} - \gamma_{j-1}}{\gamma_j}. \tag{6.3.8}$$

(d) **Epoch age distribution:** Fix $1 \leq j \leq J$. Generate $X_{PE}^{j-1}$ as above. Conditional on $X_{PE}^{j-1} = k$, generate random variable $\mathsf{Epoch}_j$ supported on the interval $[0, \alpha_j]$ with distribution

$$\mathbb{P}(\mathsf{Epoch}_j > u) := \frac{\int_0^{\alpha_j - u} m_{f_j}(\alpha_j - u - s) d\mu_{f_j}^{(k)}(ds)}{\int_0^{\alpha_j} m_{f_j}(\alpha_j - s) \mu_{f_j}^{(k)}(ds)}, \qquad 0 \leq u \leq \alpha_j. \tag{6.3.9}$$

(e) **Alive after epoch degree distribution:** Conditional on the random variables in (d) let $X_{AE}^j := \xi_{f_j}[0, \mathsf{Epoch}_j]$ where as before $\xi_{f_j}$ is the point process with attachment function $f_j$.

(f) **Alive before epoch distribution:** Fix $j \geq 1$. For $k \geq 0$, let $\xi_{f_j}^{(k)}$ be the point process (6.2.6) using attachment function $f_j$. Generate $X_{PE}^{j-1}$ as in (b). Conditional on $X_{PE}^{j-1} = k$, let $\mathfrak{C}_j := \xi_{f_j}^{(k)}[0, \alpha_j]$ with $\alpha_j$ as in (6.3.8). Define the random variable $X_{BE}^j := X_{PE}^{j-1} + \mathfrak{C}_j$.

(g) **Mixture distribution:** Finally define $X_{PE}^j$ as the following mixture: with probability $\gamma_j / \gamma_{j+1}$ $X_{PE}^j = X_{BE}^j$; with probability $(\gamma_{j+1} - \gamma_j)/\gamma_{j+1}$, let $X_{PE}^j = X_{AE}^j$.

(h) Let $\mathbf{p}^j$ be the probability mass function of $X_{PE}^j$.

With $\boldsymbol{\theta} := (\boldsymbol{\gamma}, \boldsymbol{f})$, write $D_{\boldsymbol{\theta}} := X^J_{PE}$.

**Theorem 6.3.13** (Standard model, multiple change points)**.** *As in Section 6.2, fix $k \geq 0$ and let $D_n(k)$ denote the number of vertices with out-degree $k$ in the tree $\mathcal{T}_n^{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}$ as above. Under Assumption 6.2.1 on the attachment functions $\boldsymbol{f}$ we have that*

$$\frac{D_n(k)}{n} \xrightarrow{P} \mathbb{P}(D_{\boldsymbol{\theta}} = k).$$

*Further the assertions of Corollaries 6.3.11 and 6.3.12 continue to hold in this regime.*

### 6.3.4 The quick big bang model

Now we consider the case where the change point happens "early" in the evolution of the process, where the change point scales like $o(n)$. To simplify notation, we specialize to the case $J = 1$, however our methodology is easily extendable to the general regime. Let $\{p^1_k : k \geq 0\}$ denote the probability mass function as in (6.3.1) but using the function $f_1$ to construct $\lambda^*$ in (6.2.3) and then $f_1$ in place of $f_0$ in (6.3.1).

Define for $\alpha > 0$ and any non-negative measure $\mu$,

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

We will work under the following assumption.

**Assumption 6.3.14.** $\mathbb{E}\left(\hat{\xi}_f(\lambda^*) \left|\log\left(\hat{\xi}_f(\lambda^*)\right)\right|\right) < \infty$.

**Remark 6.** *Assumption 6.3.14 is weaker than Assumption 6.3.2 as seen by considering the linear preferential attachment model with attachment function $f(i) = i + 1,\ i \geq 0$. In this case, $E\left(\hat{\xi}_f(\lambda^*)\right)^2 = \infty$ but $E\left(\hat{\xi}_f(\lambda^*)\right)^\beta < \infty$ for any $1 < \beta < 2$ (see (Bhamidi, 2007, Proposition 53 (a))).*

Recall that in the previous section, one of the messages was that the initializer function $f_0$ determined various macroscopic properties of the degree distribution for the standard model.

**Theorem 6.3.15.** *Suppose $\tau_1 = n^\gamma$ for fixed $0 < \gamma < 1$. If $f_0, f_1$ satisfy Assumptions 6.2.1, 6.3.1 and 6.3.14, the degree distribution* **does not** *feel the effect of the change point or the initializer attachment function $f_0$*

*in the sense that for any fixed $k \geq 0$,*

$$\frac{D_n(k)}{n} \xrightarrow{P} p_k^1, \qquad \text{as } n \to \infty.$$

**Remark 7.** *The form $\tau_1 := n^\gamma$ was assumed for simplicity. We believe the proof techniques are robust enough to handle any $\tau_1 = \omega_n$, where $\omega_n = o(n)$ and $\omega_n \uparrow \infty$. We defer this to future work.*

The next result implies that the maximal degree does feel the effect of the change point. Instead of proving a general result we will consider the following special cases. Throughout $M_n(1)$ denotes the maximal degree in $\mathcal{T}_n^{\boldsymbol{\theta}}$.

**Theorem 6.3.16** (Maximal degree under quick big bang)**.** *Once again assume $\tau_1 = n^\gamma$. Consider the following special cases:*

(a) **Uniform $\rightsquigarrow$ Linear:** *Suppose $f_0 \equiv 1$ whilst $f_1(k) = k + 1 + \alpha$ for fixed $\alpha > 0$. Then with high probability as $n \to \infty$, for any sequence $\omega_n \uparrow \infty$,*

$$\frac{n^{\frac{1-\gamma}{2+\alpha}} \log n}{\omega_n} \ll M_n(1) \ll n^{\frac{1-\gamma}{2+\alpha}} (\log n)^2.$$

(b) **Linear $\rightsquigarrow$ Uniform:** *Suppose $f_0(k) = k+1+\alpha$ for fixed $\alpha > 0$ whilst $f_1(\cdot) \equiv 1$. Then with high probability as $n \to \infty$, for any sequence $\omega_n \uparrow \infty$,*

$$\frac{n^{\frac{\gamma}{2+\alpha}} \log n}{\omega_n} \ll M_n(1) \ll n^{\frac{\gamma}{2+\alpha}} (\log n)^2.$$

(c) **Linear $\rightsquigarrow$ Linear:** *Suppose $f_0(k) = k + 1 + \alpha$ whilst $f_1(k) = k + 1 + \beta$ where $\alpha \neq \beta$. Then $M_n(1)/n^{\eta(\alpha,\beta)}$ is tight and bounded away from zero where*

$$\eta(\alpha, \beta) := \frac{\gamma(2+\beta) + (1-\gamma)(2+\alpha)}{(2+\alpha)(2+\beta)}. \tag{6.3.10}$$

**Remark 8.** *It is instructive to compare the above results to the setting without change point. For the uniform $f \equiv 1$ model, it is known (Devroye and Lu, 1995; Szymanski, 1990) that the maximal degree scales like $\log_2(n)$ whilst for the linear preferential attachment, the maximal degree scales like $n^{1/(\alpha+2)}$ (Móri, 2007). Thus for example, (b) of the above result coupled with Theorem 6.3.15 implies that the limiting*

*degree distribution in this case is the same as that of the uniform random recursive tree (URRT) namely Geometric with paratemer* $1/2$*; however the maximal degree scales polynomially in* $n$ *and* **not** *like* $\log n$ *as in the URRT.*

**Remark 9.** *For any* $\tau_1 \to \infty$*, the initial segment should always leave its signature in some functional of the process. See for example (Bubeck et al., 2015, 2017; Curien et al., 2014) where the evolution of the system (using typically linear preferential attachment albeit (Bubeck et al., 2017) also considered the uniform attachment case) starting from a fixed "seed" tree was considered and the aim was to detect (upto some level of accuracy) this seed tree after observing the tree* $\mathcal{T}_n$*. Similar heuristics suggest that in the context of our model, the initial segment of the process however small should show its signature at some level. We discuss this aspect further in Section 6.4.*

Proofs of results for the quick big bang model are given in Section 6.8.

### 6.3.5 Change point detection

In this Section, we discuss the statistical issues of actual change point detection from an observation of the network. We will only consider the standard model and one change point ($J = 1$). We do not believe the estimator below is "optimal" in terms of rates of convergence, however the motivation behind proving the sup-norm convergence result Theorem 6.3.6 is to provide impetus for further research in obtaining better estimators.

Consider any two sequences $h_n \to \infty, b_n \to \infty$ satisfying $\frac{\log h_n}{\log n} \to 0$, $\frac{\log b_n}{\log n} \to 0$ as $n \to \infty$. We define the following change point estimator:

$$\hat{T}_n = \inf\left\{ t \geq \frac{1}{h_n} : \sum_{k=0}^{\infty} 2^{-k} \left| \frac{D_n(k, \mathcal{T}^{\boldsymbol{\theta}}_{\lfloor nt \rfloor})}{nt} - \frac{D_n(k, \mathcal{T}^{\boldsymbol{\theta}}_{\lfloor n/h_n \rfloor})}{n/h_n} \right| > \frac{1}{b_n} \right\}.$$

The following theorem establishes the consistency of the above estimator.

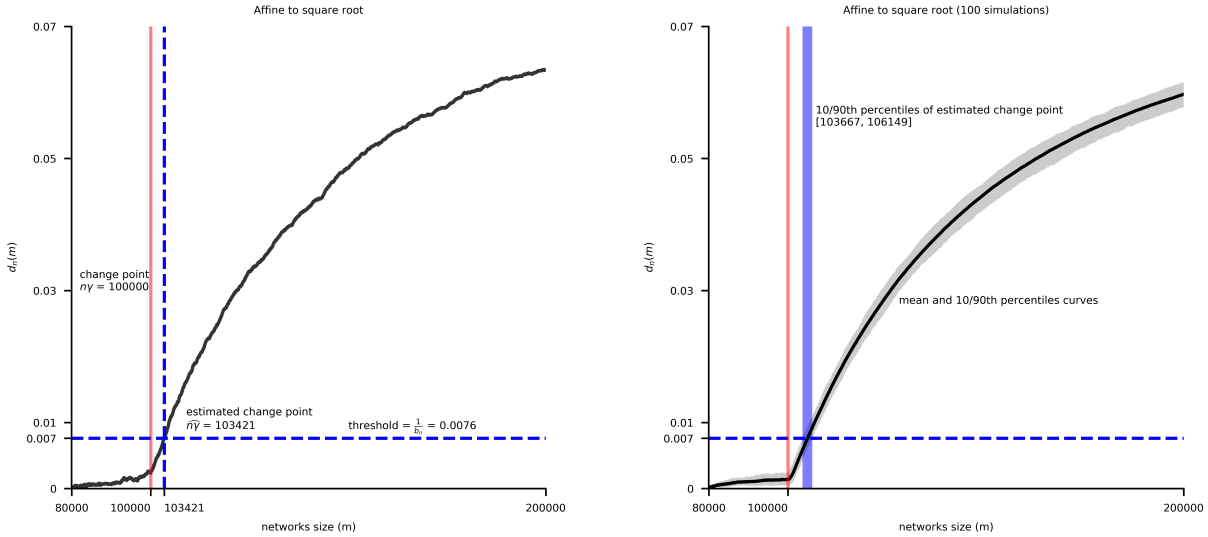
**Theorem 6.3.17.** *Assume that* $\mathbf{p}^0 \neq \mathbf{p}^1$*. Suppose* $f_0$ *satisfies Assumptions 6.2.1, 6.3.1 and 6.3.2, and* $f_1$ *satisfies Assumption 6.2.1. Then* $\hat{T}_n \xrightarrow{\mathrm{P}} \gamma$*.*

**Remark 10.** *From a practical point of view, for the proposed estimator to be close to the change point even for moderately large* $n$*, we should select* $h_n, b_n$ *satisfying the above hypotheses so that* $h_n$ *grows as slowly as possible (which ensures that we look at the evolving tree not too early, before the 'law of large numbers'*

*effect has set in) and $b_n$ grows as quickly as possible (to ensure that the detection threshold is sufficiently close to zero to capture the change in degree distribution close to the change point). One reasonable choice is $h_n = \log\log n$ and $b_n = n^{1/\log\log n}$.*

Theorem 6.3.17 is proved in Section 6.10. Figure 6.3.1 shows the result of computing the change point estimator for a network with a single change point. We plot the function:

$$d_n(m) := \sum_{k=0}^{\infty} 2^{-k} \left| \frac{D_n(k, \mathcal{T}_m^{\boldsymbol{\theta}})}{m} - \frac{D_n(k, \mathcal{T}_{\lfloor n/h_n \rfloor}^{\boldsymbol{\theta}})}{n/h_n} \right|, \qquad \frac{n}{\log\log n} < m.$$




**(a)** $d_n(\cdot)$ for one network. **(b)** Mean, $10/90th$ percentiles from 100 simulations.

**Figure 6.3.1:** The function $d_n(\cdot)$. Here $n = 2 * 10^5$, $\gamma = 0.5$, $f_0(i) = i + 2$, $f_1(i) = \sqrt{i+2}$, $h_n = \log\log n$, $b_n = n^{1/\log\log n}$. (A) The vertical, red line shows the true change point. The vertical, blue, dashed line shows the estimated change point. The horizontal, dashed, blue line shows the threshold value $b_n$. (B) The black curve shows the mean of $d_n(\cdot)$ and the grey, curved region shows the $10/90th$ percentiles (computed from 100 simulations). The blue, vertical region shows $10/90th$ percentiles of the estimated change point.

## 6.4 Discussion

(i) **Random recursive trees:** Random recursive trees have now been studied for decades, motivated by a wide array of fields including convex hull algorithms, linguistics, epidemiology and first passage percolation and recently in the study of various coalescent processes. See (Mahmoud, 2008; Drmota, 2009; Smythe and Mahmoud, 1995; Devroye, 1998; Goldschmidt and Martin, 2005) and the

references therein for starting points to this vast literature. For specific examples such as the uniform attachment or the linear attachment model with $f(i) := i+1$, one can use the seminal work of Janson (Janson, 2004) via a so-called "super ball" argument to obtain functional central limit theorems for the degree distribution. Obtaining quantitative error bounds let alone weak convergence results in the general setting considered in this chapter is much more non-trivial. Regarding proof techniques, we proceed via embedding the discrete time models into continuous time branching processes and then using martingale/renewal theory arguments for the corresponding continuous time objects to read off corresponding results for the discrete models; this approach goes back all the way to (Athreya and Karlin, 1968). Limit results for the corresponding CTBPs in the setting of interest for this chapter were developed in the seminal work of Jagers and Nerman (Jagers, 1975; Jagers and Nerman, 1984b; Nerman, 1981). One contribution of this work is to derive quantitative versions for this convergence, a topic less explored but required to answer questions regarding statistical estimation of the change point.

(ii) **Fringe convergence of random trees:** A second aim of this work (albeit not developed owing to space) is understanding rates of convergence of the *fringe distribution*. We briefly describe the context, referring the interested reader to (Aldous, 1991; Holmgren et al., 2017) for general theory and discussion of their importance in computer science. Let $\mathbb{T}$ denote the space of all rooted (unlabelled) finite trees (with $\emptyset$ denoting the empty tree). Fix a finite non-empty rooted tree $\mathcal{T} \in \mathbb{T}$ with root $\rho$. For each $v \in \mathcal{T}$ let $f(v,\mathcal{T})$ denote the sub-tree consisting of the set of vertices "below" $v$ namely vertices for which the shortest path from $\rho$ needs to pass through $v$. View $f(v,\mathcal{T})$ as an element in $\mathbb{T}$ via rooting it at $v$. The *fringe* distribution of $\mathcal{T}$ is the probability distribution on $\mathbb{T}$:

$$\pi_{\mathcal{T}}(\mathbf{t}) := \frac{1}{|\mathcal{T}|} \sum_{v \in \mathcal{T}} \mathbb{1}\left\{f(v,\mathcal{T}) = \mathbf{t}\right\}, \qquad \mathbf{t} \in \mathbb{T}.$$

If $\{\mathcal{T}_n : n \geq 1\}$ is a sequence of *random* trees, one now obtains a sequence of random probability measures. Aldous in (Aldous, 1991) shows that convergence of the associated fringe measures implies convergence of the associated random trees *locally* to limiting infinite random trees with a single infinite path; this then implies convergence of a host of global functionals such as the empirical spectral distribution of the adjacency matrix, see e.g. (Bhamidi et al., 2012). For a number of discrete random tree models, embedding these in continuous time models and using results of (Jagers

and Nerman, 1984b; Nerman, 1981) has implied convergence of this fringe distribution; however establishing rates of convergence has been non-trivial (Holmgren et al., 2017). While many of the results in this chapter are all formulated in terms of the degree distribution, the results and most of the proofs in Section 6.9 extend to more general characteristics such as the fringe distribution. To keep the paper to manageable length, this is deferred to future work.

(iii) **General change point:** Change point detection especially in the context of univariate time series has also matured into a vast field, see (Csörgö and Horváth, 1997; Brodsky and Darkhovsky, 2013). Even in this context, consistent estimation especially in the setting of multiple change points is non-trivial and requires specific assumptions on the nature of the change see e.g. (Yao, 1988) for work in estimating the change in mean of a sequence of independent observations from the normal distribution; in the context of econometric time series settings including linear regression see for example (Bai, 1997; Bai and Perron, 1998, 2003); for recent applications in the biological sciences (Olshen et al., 2004; Zhang and Siegmund, 2007). The only pre-existing work on change point in the context of evolving networks formulated in this chapter that we are aware of was carried out in (Bhamidi et al., 2015) where one assumed linear attachment functionals of the form $f(k) = k + \alpha$ for some parameter $\alpha \geq 0$. In this context, specialized computations specific to this model enabled one to derive change point detection estimators that were $\log n/\sqrt{n}$ consistent. Unfortunately these techniques do not extend to the general case considered in this chapter.

(iv) **Open questions:** In the context of rates of convergence, one natural question is to understand if one can obtain tighter bounds than those in Theorem 6.3.3 and in particular prove a functional central limit theorem (FCLT) with $\sqrt{n}$ scaling as in (Janson, 2004). In fact, a more general FCLT for the model with change point of the following form should hold: there exists a Gaussian process $\{G_\infty(t)\}_{t\in[0,1]}$ such that for any $\epsilon \in (0,1)$,

$$\left\{\sqrt{n}\left(\sum_{k=0}^{\infty} 2^{-k}\left|\frac{D_n(k,nt)}{nt} - (\Phi_{a_t}(\mathbf{p^0}))_k\right|\right)\right\}_{t\in[\epsilon,1]} \xrightarrow{d} \{G_\infty(t)\}_{t\in[\epsilon,1]}$$

in $C[\epsilon,1]$. This will directly imply $\log n/\sqrt{n}$ consistency for the proposed change point estimator. One of the major issues that this chapter does not address is the question of consistently estimating multiple change points. In the context of univariate change point detection, one is able to often use

methodology for estimating a single change point to sequentially estimate multiple change points. However the non-ergodic nature of evolution of the model considered in this chapter after the first change point does not lend itself easily to this scheme of analysis. A second line of work that we are currently exploring is extending the above techniques to general network (i.e. non-tree) models.

## 6.5 Initial embeddings and constructions

The rest of the paper is devoted to proofs of the main results.

### 6.5.1 Road map for proofs of the main results

The rest of this section is devoted to some preliminary estimates and constructions that will then be repeatedly used in the proofs. Although the results about convergence rates for the model without change point are stated before the change point results, the proof of Theorem 6.3.3 is quite technical and an essential ingredient is a “sup-norm estimate" given in Lemma 6.7.11 which is proven more generally in the context of a change point. Thus, we defer the proof of Theorem 6.3.3 to Section 6.9. Section 6.6 deals with a continuous time version of the change point model analyzed for a fixed time $a$ after the change point. Theorem 6.6.1 proved here estimates for a general characteristic $\phi \in \mathscr{C}$ the $L^1$-error in approximating the aggregate $\phi$-score at time $a$ of all individuals born after the change point with a weighted linear combination of the degree counts at the change point. This estimate, apart from directly yielding a law of large numbers (see second part of Theorem 6.6.1), turns out to be crucial in most subsequent proofs. The estimates derived in Section 6.6 are then used in Section 6.7 to analyze the standard model and prove the main theorems in this setting (Theorems 6.3.6 and 6.3.9) as well as Corollary 6.3.11 on the initializer always winning. Corollary 6.3.10 follows directly from Lemma 6.10.3 and requires an in-depth analysis of the fluid limits derived in Theorem 6.3.6 and is postponed to Section 6.10. Section 6.8 contains proofs of the quick big bang model. We note here that all the estimates obtained in Sections 6.6 and 6.7 to analyze the model for a fixed time $a$ after the change point explicitly exhibited the dependence on $a$. This turns out to be crucial in Section 6.8 where we take $a = \eta_0 \log n$ and the estimates above still hold if $\eta_0$ is sufficiently small. Roughly speaking, we partition the interval $[T_{n^\gamma}, T_n]$ into finitely many subintervals of size at most $\eta_0 \log n$ and ‘bootstrap’ the estimates obtained above to prove Theorem 6.3.15. We conclude in Section 6.10 with the proof of Theorem 6.3.17 on the change point detection estimator.

### 6.5.2 Initial constructions

Fix $n \geq 3$, and $1 < r_n < n$ ($r_n$ will later assume the value $\gamma n$ or $n^\gamma$), two attachment functions $f_0, f_1$ satisfying Assumption 6.2.1.

**Definition 6.5.1** ( CTBP with change point)**.** *Recall that* $\{\mathrm{BP}_{f_0}(t) : t \geq 0\}$ *denotes a continuous-time branching process driven by the point process* $\xi_{f_0}$ *defined in* (6.2.4)*. Now for* $n, r_n$ *and two attachment functions* $f_0, f_1$ *as above define* $\{\mathrm{BP}_n(t) : t \geq 0\}$ *as follows:*

(a) *Generate a process* $\mathrm{BP}_{f_0}(\cdot)$ *as above. For* $0 \leq t \leq T_{r_n}$ *let* $\mathrm{BP}_n(\cdot) \equiv \mathrm{BP}_{f_0}(\cdot)$.

(b) *At time* $T_{r_n}$ *all existing vertices change their reproduction, so that for any fixed* $k \geq 0$*, a vertex with* $k$ *children in* $\mathrm{BP}_{f_0}(T_{r_n})$ *now uses offspring distribution* $\xi_{f_1}^{(k)}$ *for all subsequent offspring. Each new vertex born into the system has offspring point process with distribution* $\xi_{f_1}$*, independent across vertices. Label vertices as above according to the time order they enter the system.*

The following is the analog of Lemma 6.2.3 in the change point setting.

**Lemma 6.5.2.** *Let* $\boldsymbol{\theta} = (f_0, f_1, r_n)$ *and consider the sequence of random trees* $\{\mathscr{T}_m^{\boldsymbol{\theta}} : 2 \leq m \leq n\}$ *with one change point at* $\tau_1 = r_n$*. Consider the continuous time construction in Definition 6.5.1 and define for* $m \geq 1$ *the stopping times* $T_m := \inf\{t \geq 0 : |\mathrm{BP}_n(t)| = m\}$*. Then viewed as a sequence of growing random labelled rooted trees we have,* $\{\mathrm{BP}_n(T_m) : 2 \leq m \leq n\} \stackrel{d}{=} \{\mathscr{T}_m^{\boldsymbol{\theta}} : 2 \leq m \leq n\}$.

The next few Lemmas deal with properties of one important class of offspring point processes that arise in the study of *linear* preferential attachment.

**Definition 6.5.3** (Rate $\nu$ Yule process)**.** *Fix* $\nu > 0$*. A rate* $\nu$ *Yule process is a pure birth process* $\{Y_\nu(t) : t \geq 0\}$ *with* $Y_\nu(0) = 1$ *and where the rate of birth of new individuals is proportional to size of the current population. More precisely*

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

*where* $\{\mathscr{F}(t) : t \geq 0\}$ *is the natural filtration of the process.*

The following is a standard property of the Yule process, see e.g. (Norris, 1998, Section 2.5).

**Lemma 6.5.4.** *Fix $t > 0$ and rate $\nu > 0$. Then $Y_\nu(t)$ has a Geometric distribution with parameter $p = e^{-\nu t}$. Precisely,*

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

*The process $\{Y_\nu(t)\exp(-\nu t) : t \geq 0\}$ is an $\mathbb{L}^2$ bounded martingale and thus there exists a strictly positive random variable $W$ such that $Y_\nu(t)\exp(-\nu t) \xrightarrow{\text{a.e.}} W$. Further $W = \exp(1)$.*

Next we derive moment bounds for the attachment point processes for linear preferential attachment.

**Lemma 6.5.5.** *Fix $\nu > 0$, $\kappa \geq 0$. Let $\xi_{\nu,\kappa}(t)$ be the offspring distribution of a linear preferential attachment process with with attachment function $f(i) = \nu(i+1) + \kappa$. Then with respect to the natural filtration the following two processes are martingales:*

$$M_1(t) := e^{-\nu t}\xi_{\nu,\kappa}(t) - \frac{\nu + k}{\nu}\left(e^{\nu t} - 1\right), \ t \geq 0$$

*and*

$$M_2(t) := e^{-2\nu t}\xi_{\nu,\kappa}(t)^2 - \int_0^t (2\kappa + 3\nu)\xi_{\nu,\kappa}(s)e^{-2\nu s}ds - \frac{\nu + \kappa}{2\nu}\left(1 - e^{-2\nu t}\right), \ t \geq 0.$$

*In particular,*

$$\mathbb{E}\,\xi_{\nu,\kappa}(t) = \frac{\nu + \kappa}{\nu}\left(e^{\nu t} - 1\right), \text{ and } \mathbb{E}\left(\xi_{\nu,\kappa}(t)\right)^2 = \frac{(2\kappa + 3\nu)(\nu + \kappa)}{2\nu^2}\left(e^{\nu t} - 1\right)^2 + \frac{\nu + \kappa}{2\nu}\left(e^{2\nu t} - 1\right).$$

*Proof.* We sketch the proof. Let $\mathscr{F}(t)$ be the natural filtration corresponding to the continuous time branching process with attachment function $f$. Note that $\xi_{\nu,\kappa}(t) \rightsquigarrow \xi_{\nu,\kappa}(t) + 1$ at rate $\nu(\xi_{\nu,\kappa}(t) + 1) + \kappa$. This can be used to check $\mathbb{E}[dM_1(t)|\mathscr{F}(t)] = 0$ showing $M_1(t)$ is a martingale. Similarly, $\xi_{\nu,\kappa}(t)^2 \rightsquigarrow \xi_{\nu,\kappa}(t)^2 + 2\xi_{\nu,\kappa}(t) + 1$ at rate $\nu(\xi_{\nu,\kappa}(t) + 1) + \kappa$. This expression can similarly be used to check $M_2(t)$ is a martingale. The first expectation claimed in the lemma follows immediately by setting the expectation of $M_1(t)$ equal to zero. The second expectation follows by computing the expectation of $M_2(t)$ and then using the expectation of $\xi_{\nu,\kappa}(t)$. ■

The next result derives moment bounds for a particular class of CTBP.

**Definition 6.5.6** (Rate $\nu$ Affine $\kappa$ PA model)**.** *Fix $\nu > 0, \kappa \geq 0$. A branching process whose offspring distribution is given by an offspring distribution constructed using attachment function $f(i) = \nu(i+1) + \kappa$ will be called a linear PA branching process with rate $\nu$ and affine parameter $\kappa$. Denote this as $\{\mathrm{PA}_{\nu,\kappa}(t) : t \geq 0\}$.*

We will now derive expressions for moments of the process $\mathrm{PA}_{\nu,\kappa}$ that will be useful in the sequel. To simplify notation, when possible we will suppress dependence on $\nu, \kappa$ and write the above as $\mathrm{PA}(\cdot)$.

**Proposition 6.5.7.** *Fix $\nu > 0, \kappa \geq 0$. With respect to the natural filtration, the following processes are Martingales:*

$$M_1(t) := e^{-(2\nu+\kappa)t}(|\mathrm{PA}_{\nu,\kappa}(t)| - 1) - \frac{\nu+\kappa}{2\nu+\kappa}(1 - e^{-(2\nu+\kappa)t}), \qquad t \geq 0$$

*and*

$$M_2(t) := (|\mathrm{PA}_{\nu,\kappa}(t)| - 1)^2 - \int_0^t ((4\nu+2\kappa)(|\mathrm{PA}_{\nu,\kappa}(s)| - 1)^2 + (4\nu+3\kappa)(|\mathrm{PA}_{\nu,\kappa}(s)| - 1) + (\nu+\kappa))ds, \quad t \geq 0.$$

*In particular, for any fixed $a > 0$, $\exists\, C$ (dependent on $\nu$ and $\kappa$ but not on $a$) such that for $0 \leq t \leq a$*

$$\mathbb{E}(|\mathrm{PA}_{\nu,\kappa}(t)|) - 1 \leq Ce^{(2\nu+\kappa)a}t; \qquad \mathbb{E}((|\mathrm{PA}_{\nu,\kappa}(t)| - 1)^2) \leq Ce^{(4\nu+2\kappa)a}t. \tag{6.5.1}$$

*Proof.* Write $\{\mathscr{F}(t) | t \geq 0\}$ for the natural filtration of the process. Note that $|\mathrm{PA}(t)| \rightsquigarrow |\mathrm{PA}(t)| + 1$ at rate $\sum_{x \in \mathrm{PA}(t)}(\nu(d_x(t)+1)+\kappa) = (2\nu+\kappa)|\mathrm{PA}(t)| - \nu$ where $d_x(t)$ is the number of children of $x$ at time $t$. This can be used to check $\mathbb{E}(dM_1(t)|\mathscr{F}(t)) = 0$. Computing expectations gives $\mathbb{E}\,\mathrm{PA}(t) - 1 = \frac{\nu+\kappa}{2\nu+\kappa}\left(e^{(2\nu+\kappa)t} - 1\right)$ from which the first moment bound follows for $t \leq a$.

Similarly, $\mathrm{PA}(t) - 1$ undergoes the change $(\mathrm{PA}(t+) - 1)^2 - (\mathrm{PA}(t) - 1)^2 = 2(\mathrm{PA}(t) - 1) + 1$ at rate $(2\nu + \kappa)(\mathrm{PA}(t) - 1) + \nu + \kappa$. This can be used to check $M_2(\cdot)$ is a martingale. Computing the expectation of this martingale gives the second moment bound. ∎

The next result which follows from (Nerman, 1981; Jagers and Nerman, 1984a) describes limit results for a number of important characteristics of relevance in this chapter. Recall the class of characteristics $\mathscr{C}$ defined in (6.3.2). Recall that $\lambda^*$ was the Malthusian rate of growth and $\mu_f$ denoted the mean measure of the offspring distribution. Let $m^\star := \int_{\mathbb{R}_+} ue^{-\lambda^* u}\mu_f(du)$. For any fixed characteristic $\chi \in \mathscr{C}$ and any $\alpha > 0$, define,

$$\hat{\chi}(\alpha) := \int_0^\infty \alpha e^{-\alpha t}\chi(t)dt.$$

Also recall for $\alpha > 0$,

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

A useful fact is that for any $\alpha > 0$, recalling $\hat{\rho}$ from Assumption 6.2.1 (iii),

$$\hat{\rho}(\alpha) = \hat{\mu}_f(\alpha) = \int_0^\infty e^{-\alpha t} \mu_f(dt).$$

Recall $Z_f^\chi(t) = \sum_{x \in \mathrm{BP}_f(t)} \chi(t - \sigma_x)$ and $M_f^\chi(t) = \mathbb{E}\left(e^{-\lambda^* t} Z_f^\chi(t)\right)$. Recall $Z_f(t)$ is the total number of vertices at time $t$ and $M_f(t) = \mathbb{E}\left(e^{-\lambda^* t} Z_f(t)\right)$. The following Lemma is a consequence of (Nerman, 1981, Theorem 6.3).

**Lemma 6.5.8.** *(i) Under Assumption 6.2.1 (iii), for any characteristic $\chi \in \mathscr{C}$,*

$$\frac{Z_f^\chi(t)}{Z_f(t)} \xrightarrow{a.s.} \mathbb{E}(\hat{\chi}(\lambda^*)).$$

*(ii) Under Assumptions 6.2.1 and 6.3.14, there exists a strictly positive random variable $W_\infty$ with $\mathbb{E}(W_\infty) = 1$ such that for characteristics $\chi \in \mathscr{C}$,*

$$e^{-\lambda^* t} Z_f^\chi(t) \xrightarrow{a.s., \mathbb{L}^1} \frac{\mathbb{E}(\hat{\chi}(\lambda^*))}{\lambda^* m^\star} W_\infty.$$

*Proof.* (i) We will apply (Nerman, 1981, Theorem 6.3) with characteristics $\chi$ and $\psi$ defined by $\psi(t) := \mathbb{1}\{t \geq 0\}$ by verifying Conditions 6.1 and 6.2 in (Nerman, 1981). Condition 6.1 holds for $\xi_f$ by Assumption 6.2.1 (iii). Condition 6.2 requires there exist $\beta < \lambda^*$ such that $\mathbb{E}\left[\sup_t \left(e^{-\beta t}\chi(t)\right)\right] < \infty$ and $\mathbb{E}\left[\sup_t \left(e^{-\beta t}\psi(t)\right)\right] < \infty$. For $\psi$ this condition holds for any $\beta$ since $\mathbb{E}\left[\sup_t \left(e^{-\beta t}\psi(t)\right)\right] = \mathbb{E}\left[\sup_t \left(e^{-\beta t}\right)\right] = 1$. To verify Condition 6.2 for $\chi$ note that for any $\beta \geq 0$, using the fact that $\chi \in \mathscr{C}$,

$$\sup_{t \in [0,\infty)} \left(e^{-\beta t}\chi(t)\right) \leq \sum_{j=0}^\infty \sup_{t \in [j, j+1)} \left(e^{-\beta t}\chi(t)\right) \leq b_\chi \sum_{j=0}^\infty e^{-\beta j}(\xi_f(j+1)+1).$$

From Assumption 6.2.1 (iii), there exists $\beta_0 < \lambda^*$ such that $\hat{\mu_f}(\beta_0) < \infty$ which implies there exists $C > 0$ such that $\mathbb{E}\left(\xi_f(t+1)+1\right) \leq Ce^{\beta_0 t}$ for all $t \geq 0$. Using this and setting $\beta = (\beta_0 + \lambda^*)/2$, we get

$$\mathbb{E}\left(\sup_{t \in [0,\infty)} \left(e^{-\beta t}\chi(t)\right)\right) \leq C b_\chi \sum_{j=0}^\infty e^{-\beta j} e^{\beta_0 j} < \infty \tag{6.5.2}$$

which verifies Condition 6.2 for $\chi$. It is easy to check condition (2.6) in (Nerman, 1981) using the fact that $\mathbb{E}\big(\chi(t)\big) \leq Cb_\chi e^{\beta_0 t}$ and $\psi(t) \leq 1$. Thus, Proposition 2.2 of (Nerman, 1981) implies $M_f^\chi(\infty) = \big(\mathbb{E}(\hat{\chi}(\lambda^*)\big)/(\lambda^* m^\star)$ and $M_f(\infty) = 1/(\lambda^* m^\star)$ and this, along with Theorem 6.3 from (Nerman, 1981) implies (i).

(ii) To show almost sure convergence, we will verify Conditions 5.1 and 5.2 of (Nerman, 1981). From (6.2.2), we obtain $\beta_0 < \lambda^*$ such that $\int_0^\infty e^{-\beta_0 t}\mu(dt) < \infty$ and this implies Condition 5.1 with $g(t) = e^{-(\lambda^*-\beta_0)t}$ (see the remark following Condition 5.1 of (Nerman, 1981)). Condition 5.2 for $\chi \in \mathscr{C}$ follows from (6.5.2) with $h(t) = e^{-(\lambda^*-\beta)t}$. The almost sure convergence now follows from Theorem 5.4 of (Nerman, 1981). The $\mathbb{L}^1$ convergence follows from Corollary 3.3 of (Nerman, 1981) upon using Assumption 6.3.14 and noting that $\mathbb{E}\big(\chi(t)\big)$ is continuous a.e. with respect to Lebesgue measure by Lemma 5.3 of (Nerman, 1981), along with a straightforward verification of conditions (3.1) and (3.2) in Theorem 3.1 of (Nerman, 1981). The positivity of $W_\infty$ follows from Proposition 1.1 of (Nerman, 1981) upon observing that the number of vertices born by time $t$ goes to infinity almost surely as $t \to \infty$. ■

### 6.6 Change point model for fixed time $a$: point-wise convergence for general characteristics

In this section we consider growing the tree for a constant time $a$ after the change point i.e. for $t \in [T_{\gamma n}, T_{\gamma n} + a]$ using the second attachment function, $f_1$. Consider the class of characteristics $\mathscr{C}$ defined in (6.3.2). We will count vertices *born after the change point* according to a general characteristic $\phi \in \mathscr{C}$ and prove a law of large numbers for this aggregate $\phi$-score at time $a$ as $n \to \infty$ (see Theorem 6.6.1). This will be a key tool in the rest of the paper. For notational convenience we will consider the time to start at $t = 0$ (i.e. $t = s$ corresponds to actual time $T_{\gamma n} + s$ for any $s \in [0, a]$). For $t \geq 0$, $\mathrm{BP}_n(t)$ will denote the branching process at time $t$ (i.e. time $t$ after the change point).

#### 6.6.1 Notation

Let $\lambda_i^*$ denote the Malthusian parameter for the branching process with attachment function $f_i$. For the branching process (without change point) with attachment function $f_1$, and for any characteristic $\phi$, recall $Z_{f_1}^\phi(t) = \sum_{x \in \mathrm{BP}_{f_1}(t)} \phi_x(t - \sigma_x)$. When $\phi(t) = \mathbb{1}\{t \geq 0\}$, we will write $Z_{f_1}$ for $Z_{f_1}^\phi$. Let $m_{f_1}^\phi(t) := \mathbb{E}\, Z_{f_1}^\phi(t)$ and let $v_{f_1}^\phi(t) = \mathrm{Var}\Big(Z_{f_1}^\phi(t)\Big)$.

For $\phi \in \mathscr{C}$, an easy computation implies there exists $c > 0$ such that $Z^{\phi}_{f_1}(t) \le 2cZ_{f_1}(t)$ for every $t \ge 0$ and hence,

$$\sup_{t\in[0,a]} m^{\phi}_{f_1}(t) \le 2c\mathbb{E}(Z_{f_1}(a)) \le Ce^{C'a}, \quad \sup_{t\in[0,a]} v^{\phi}_{f_1}(t) \le 4c^2\mathbb{E}(Z^2_{f_1}(a)) \le Ce^{C'a} \tag{6.6.1}$$

where $C, C'$ do not depend on $a$. This follows by Assumption 6.2.1(ii) on $f_1$ that implies $\mathrm{BP}_{f_1}(\cdot)$ is stochastically dominated by a rate $C$ PA branching processes (see Definition 6.5.6) and then by appealing to (6.5.1).

### 6.6.2 Definitions

Next we define various constructs which will be used in this section. Divide the interval $[0,a] := \cup_{i=0}^{n^\delta-1}[ia/n^\delta, ((i+1)a)/n^\delta]$ into subintervals of size $a/n^\delta$. We will eventually take limits as $\delta \to \infty$.

(i) **System at change point:** Recall the construction of the change point model in continuous time via Lemma 6.5.2. Let $\mathscr{F}_n(0)$ denote the $\sigma-$field containing the information till $T_{n\gamma}$, the change point. Define the filtration $\{\mathscr{F}_n(t) : t \ge 0\} := \{\sigma(\mathrm{BP}_n(t)) : t \ge 0\}$. We will first work conditional on $\mathscr{F}_n(0)$. For fixed $k \ge 0$, to specify dependence on time, we write $\mathscr{D}_n(k,t)$ to be the set of vertices with (out-)degree $k$ at time $t$ and let $D_n(k,t) := |\mathscr{D}_n(k,t)|$. The initial set $\mathscr{D}_n(k,0)$ which arose from the pre-change point dynamics will play a special role. Label the vertices in $\mathscr{D}_n(k,0)$ in the order they were born into $\mathrm{BP}_n(0)$ as $\mathscr{D}_n(k,0) := \left\{v^{(k)}_1, v^{(k)}_2, \ldots, v^{(k)}_{D_n(k,0)}\right\}$.

(ii) **Descendants in small intervals:** For $0 \le i \le n^\delta - 1$ and $v^{(k)}_j \in \mathscr{D}_n(k,0)$, we track evolution of descendants of this vertex in the various subintervals. Let $\mathscr{V}^{(k)}_n(i,j)$ denote the set of children born in the interval $\left[\frac{ia}{n^\delta}, \frac{(i+1)a}{n^\delta}\right]$ to $v^{(k)}_j$. Let $N^{(k)}_n(i,j) := |\mathscr{V}^{(k)}_n(i,j)|$ be the number of such vertices. Write $N^{(k)}_n(i) := \sum_{j=1}^{D_n(k,0)} N^{(k)}_n(i,j)$.

(iii) **Good and bad vertices:** Call a vertex in $\mathscr{V}^{(k)}_n(i,j)$ a *good* vertex if it does **not** give birth to any children by $\frac{(i+1)a}{n^\delta}$. Let $\widetilde{\mathscr{V}}^{(k)}_n(i,j) \subseteq \mathscr{V}^{(k)}_n(i,j)$ denote the set of good children born in the interval $\left[\frac{ia}{n^\delta}, \frac{(i+1)a}{n^\delta}\right]$ born to $v^{(k)}_j$. Let $\widetilde{N}^{(k)}_n(i,j) := |\widetilde{\mathscr{V}}^{(k)}_n(i,j)|$ be the number of such vertices. As above, write $\widetilde{N}^{(k)}_n(i) := \sum_{j=1}^{D_n(k,0)} \widetilde{N}^{(k)}_n(i,j)$ be the total number of *good* children born to vertices which originally had degree $k$ at the change point. Let $\mathscr{B}^{(k)}_n(i,j) := \mathscr{V}^{(k)}_n(i,j) \setminus \widetilde{\mathscr{V}}^{(k)}_n(i,j)$ be the collection of bad children namely those in $\mathscr{V}^{(k)}_n(i,j)$ who **have** reproduced by time $\frac{(i+1)a}{n^\delta}$. Let $B^{(k)}_n(i,j) = |\mathscr{B}^{(k)}_n(i,j)|$. Let $\mathscr{R}^{(k)}_n(i,j)$ be the set of descendants born in $\left[\frac{ia}{n^\delta}, \frac{(i+1)a}{n^\delta}\right]$ to vertices in $\mathscr{B}^{(k)}_n(i,j)$ and let $R^{(k)}_n(i,j) := |\mathscr{R}^{(k)}_n(i,j)|$.

(iv) **Vertices counted by a characteristic:** Let $Z_n^{(k),\phi}(i,j,x)$ be the number of descendents, satisfying characteristic $\phi$, at time $a$, born to $x \in \mathcal{V}_n^{(k)}(i,j)$. Write $Z_n^{(k),\phi} = \sum_{j=1}^{D_n(k,0)} \sum_{i=0}^{n^\delta-1} \sum_{x\in\mathcal{V}_n^{(k)}(i,j)} Z_n^{(k),\phi}(i,j,x)$. Let $Z_n^{\phi} = \sum_{k=1}^{\infty} Z_n^{(k),\phi}$. Let $\widetilde{Z}_n^{(k),\phi}$ be the number of such descendants as above, but born to a good parent i.e. let $\widetilde{Z}_n^{(k),\phi} = \sum_{j=1}^{D_n(k,0)} \sum_{i=0}^{n^\delta-1} \sum_{x\in\widetilde{\mathcal{V}}_n^{(k)}(i,j)} Z_n^{(k),\phi}(i,j,x)$. Let $\xi_{f_i}^{(k)}[s,t]$ denote the distribution of the number of children born in the time interval $[s,t]$ to a vertex who had degree $k$ at time 0 with attachment function $f_i$. Write $\xi_{f_i}^{(k)}(t)$ for $\xi_{f_i}^{(k)}[0,t]$.

(v) **Technical conditioning tool:** Define the following $\sigma-$field

$$\mathcal{G}_n = \sigma\Big(\mathcal{F}_n(0) \bigcup \{\text{life history of } v \in \mathrm{BP}_n(0) \text{ till time a}\} \bigcup_{j\le n^\delta-1} \left\{\text{all vertices born in } \left[\frac{ja}{n^\delta}, \frac{(j+1)a}{n^\delta}\right] \text{ and their life history till time } \frac{(j+1)a}{n^\delta}\right\}\Big).$$

(vi) **Mean of characteristics emanating from degree $k$ parent:** Let $\lambda_k^{\phi}(t) = \int_0^t m_{f_1}^{\phi}(t-s)\mu_{f_1}^{(k)}(ds)$ for $t \le a$. For notational simplicity since $a$ is fixed in this Section, we will write $\lambda_k^{\phi} := \lambda_k^{\phi}(a)$.

The following is the main result we prove in this section.

**Theorem 6.6.1.** *Fix any $\phi \in \mathscr{C}$. There exist deterministic positive constants $C, C' < \infty$ (not dependent on $a$) such that for every $a > 0$ and $n \ge 2$,*

$$\mathbb{E}\left[\left|Z_n^{\phi} - \sum_{k=0}^{\infty} D_n(k,0)\lambda_k^{\phi}\right| \Big| \mathcal{F}_n(0)\right] \le Ce^{C'a}\sqrt{n}.$$

*In particular, as $n \to \infty$,*

$$\frac{Z_n^{\phi}}{n} \xrightarrow{\mathrm{P}} \gamma \sum_{k=0}^{\infty} p_k^0 \lambda_k^{\phi}(a).$$

### 6.6.3 Proof of Theorem 6.6.1:

We fix a characteristic $\phi \in \mathscr{C}$ throughout the proof. The main tools in order to prove this result are Lemmas 6.6.10, 6.6.11 below. In order to prove these results we will need a number of supporting results which we now embark upon. First we start with a technical lemma controlling the number of children a vertex with degree $k$ at change point can produce within a fixed interval. For the rest of this section we write $C_1, C_2, C_3, C_4, C, C', a_0$ for constants which are independent of $a, n, \delta, k$.

**Lemma 6.6.2.** *For any interval* $[b, b+\eta] \subseteq [0, a]$,

$$\mathbb{E}\left[\xi_{f_1}^{(k)}[b, b+\eta]\right] \leq C_1 e^{C_2 a}(k+1)\eta, \qquad \mathbb{E}\left[\xi_{f_1}^{(k)}[b, b+\eta]^2\right] \leq C_3 e^{C_4 a}\left\{(k+1)^2\eta^2 + (k+1)\eta\right\}.$$

*Proof.* By Assumption 6.2.1(ii), the process $\left\{U(t) := \xi_{f_1}^{(k)}(t/C) : t \geq 0\right\}$ is stochastically dominated by the offspring distribution of a linear preferential attachment (**PA**) $\{P_k(t) : t \geq 0\}$ point process started at $k+1$, namely a point process constructed using attachment function $f^{(k)}(i) = k+1+i$ for $i \geq 0$ with initial condition $P_k(0) := 0$. From the first moment bound in Lemma 6.5.5 (with $\nu = 1$ and $\kappa = k$) we find

$$\mathbb{E}(P_k(t)) = (1+k)(e^t - 1) \tag{6.6.2}$$

We show how to use the first moment of $P_k(\cdot)$ to obtain the first assertion in the Lemma. The second assertion follows from the same argument using the second moment of $P_k(\cdot)$ which is also obtained from Lemma 6.5.5. Conditioning on $\xi_{f_1}^{(k)}(b)$ and using the Markov property we get,

$$\mathbb{E}\xi_{f_1}^{(k)}[b, b+\eta] = \sum_{d=0}^{\infty} \mathbb{P}\left(\xi_{f_1}^{(k)}(b) = d\right) \mathbb{E}\xi_{f_1}^{(k+d)}(\eta) \tag{6.6.3}$$

Now for any fixed $k \geq 0$ and $t \leq a$, using domination by the corresponding PA process, we get

$$\mathbb{E}[\xi_{f_1}^{(k)}(t)] \leq \mathbb{E}(P_k(tC)) = e^{tC}(1+k)(1-e^{-tC}) \leq e^{Ca}C(1+k)t. \tag{6.6.4}$$

Using this bound twice in (6.6.3) gives,

$$\begin{aligned}\mathbb{E}\xi_{f_1}^{(k)}[b, b+\eta] \leq Ce^{Ca}\eta \sum_{d=0}^{\infty} \mathbb{P}\left(\xi_{f_1}^{(k)}(b) = d\right)(1+k+d) = Ce^{Ca}\eta(1+k+\mathbb{E}(\xi_{f_1}^{(k)}(b))) \\ \leq Ce^{Ca}\eta(1+k+Cbe^{Ca}C(1+k)) \leq C'e^{C''a}(k+1)\eta \end{aligned} \tag{6.6.5}$$

where $C', C''$ are constants that do not depend on $k, a$. This completes the proof.

■

Recall that conditional on the initial $\sigma$-field $\mathscr{F}_n(0)$, the random variable $N_n^{(k)}(i,j) \stackrel{d}{=} \xi_{f_1}^{(k)}\left[\frac{ia}{n^\delta}, \frac{(i+1)a}{n^\delta}\right]$. Using Lemma 6.6.2 now gives the following result.

**Corollary 6.6.3.** *For all* $1 \le j \le D_n(k,0)$, $\mathbb{E}(N_n^{(k)}(i,j)|\mathscr{F}_n(0)) \le C_1 e^{C_2 a}(k+1)n^{-\delta}$ *and* $\mathbb{E}\left[N_n^{(k)}(i,j)^2|\mathscr{F}_n(0)\right] \le C_3 e^{C_4 a}\left\{(k+1)^2 n^{-2\delta} + (k+1)n^{-\delta}\right\}$.

The next Lemma bounds the number of "bad" vertices and their descendants born within small intervals. For the rest of this section, unless specified otherwise we always work conditional on $\mathscr{F}_n(0)$ so that expectation operations such as $\mathbb{E}(\cdot)$ and $\mathrm{Var}(\cdot)$ in the ensuing results mean $\mathbb{E}(\cdot|\mathscr{F}_n(0))$ and $\mathrm{Var}(\cdot|\mathscr{F}_n(0))$.

**Lemma 6.6.4.** *For any* $k,i,j$,

$$\mathbb{E}(R_n^{(k)}(i,j)) \le C_1 e^{C_2 a}\frac{(k+1)}{n^{2\delta}}, \qquad \mathbb{E}\left(\left(R_n^{(k)}(i,j)\right)^2\right) \le C_3 e^{C_4 a}\left(\frac{(k+1)}{n^{2\delta}} + \frac{(k+1)^2}{n^{4\delta}}\right).$$

*Proof.* For every child $u \in \mathcal{V}_n^{(k)}(i,j)$, write $\mathrm{BP}(\cdot;u)$ for the branching process lineage emanating from $u$. Conditional on $\mathcal{V}_n^{(k)}(i,j)$, using Assumption 6.2.1(ii) on $f_1$, generate a collection of independent rate $C$ PA branching processes (see Definition 6.5.6) $\left\{Y_\ell : 1 \le \ell \le |\mathcal{V}_n^{(k)}(i,j)|\right\}$ such that $|\mathrm{BP}(\cdot;u)| \le |Y_\ell(\cdot)|$. Now note that $X_\ell(t) := Y_\ell(t) - 1$ is the number of descendants of the root for this branching process by time $t$. Using this construction we have the trivial inequality $R_n^{(k)}(i,j) \le \sum_{\ell=1}^{N_n^{(k)}(i,j)} X_\ell\left[0, \frac{a}{n^\delta}\right]$. This implies

$$\mathbb{E}(R_n^{(k)}(i,j)) \le \mathbb{E}(N_n^{(k)}(i,j))\,\mathbb{E}\left(X_1\left[0,\frac{a}{n^\delta}\right]\right),$$

and

$$\mathbb{E}\left(\left[R_n^{(k)}(i,j)\right]^2\right) \le \mathbb{E}(N_n^{(k)}(i,j))\,\mathbb{E}\left(\left[X_1\left[0,\frac{a}{n^\delta}\right]\right]^2\right) + \mathbb{E}\left(\left[N_n^{(k)}(i,j)\right]^2\right)\left(\mathbb{E}\left(X_1\left[0,\frac{a}{n^\delta}\right]\right)\right)^2.$$

Corollary 6.6.3 for moments of $N_n^{(k)}(i,j)$ and (6.5.1) for moments of $X_1\left[0,\frac{a}{n^\delta}\right]$ completes the proof.

■

The next Lemma bounds fluctuations of *good* descendants of degree $k$ ancestors at the change point counted according to a characteristic.

**Lemma 6.6.5.** *For any* $k \ge 0$, $\mathrm{Var}\left(\widetilde{Z}_n^{(k),\phi}\right) \le Ce^{C'a}\left((k+1)^2 n^{-\delta} + (k+1)\right) D_n(k,0)$.

*Proof.* By construction we have

$$\mathrm{Var}\left(\widetilde{Z}_n^{(k),\phi}\right) = \mathrm{Var}\left(\sum_{j=1}^{D_n(k,0)}\sum_{i=0}^{n^\delta-1}\sum_{x\in\widetilde{\mathcal{V}}_n^{(k)}(i,j)} Z_n^{(k),\phi}(i,j,x)\right) = D_n(k,0)\,\mathrm{Var}\left(\sum_{i=0}^{n^\delta-1}\sum_{x\in\widetilde{\mathcal{V}}_n^{(k)}(i,1)} Z_n^{(k),\phi}(i,1,x)\right). \quad (6.6.6)$$

We analyze the variance term on the right by first conditioning on $\mathcal{G}_n$. First note that,

$$\mathbb{E}\left[\mathrm{Var}\left(\sum_{i=0}^{n^\delta-1}\sum_{x\in\widetilde{\mathcal{V}}_n^{(k)}(i,1)} Z_n^{(k),\phi}(i,1,x)\Big|\mathcal{G}_n\right)\right] = \mathbb{E}\left[\sum_{i=0}^{n^\delta-1}\widetilde{N}_n^{(k)}(i,1)\,\nu_{f_1}^{\phi}\left(a-\frac{(i+1)a}{n^\delta}\right)\right]$$
$$\leq C_1 e^{C_2 a}(k+1)n^{-\delta}n^{\delta}\,\mathbb{E}(Z_{f_1}^2(a)) \leq Ce^{C'a}(k+1) \tag{6.6.7}$$

where $C, C'$ do not depend on $k, a, n, \delta$. The first equality comes from noting $\widetilde{\mathcal{V}}_n^{(k)}(i,1)$ is $\mathcal{G}_n^{(k)}$ measurable, the collection $\left\{Z_n^{(k),\phi}(i,1,x) | x\in\widetilde{\mathcal{V}}_n^{(k)}(i,1), 1\leq i\leq n^\delta-1\right\}$ are independent and further for each $0\leq i\leq n^\delta-1$ and $x\in\widetilde{\mathcal{V}}_n^{(k)}(i,1)$ , $Z_n^{(k),\phi}(i,1,x)$ is distributed as $Z_{f_1}^{\phi}\left(a-\frac{(i+1)a}{n^\delta}\right)$, since $x$ has no children by time $\frac{(i+1)a}{n^{-\delta}}$. The second inequality follows by using Corollary 6.6.3 for $N_n^{(k)}(i,1)$ and (6.6.1). Similarly

$$\mathrm{Var}\left(\mathbb{E}\left[\sum_{i=0}^{n^\delta-1}\sum_{x\in\widetilde{\mathcal{V}}_n^{(k)}(i,1)} Z_n^{(k),\phi}(i,j,x)\Big|\mathcal{G}_n\right]\right) = \mathrm{Var}\left(\sum_{i=0}^{n^\delta-1}\widetilde{N}_n^{(k)}(i,1)m_{f_1}^{\phi}\left(a-\frac{(i+1)a}{n^\delta}\right)\right)$$
$$\leq 4c^2\left(\mathbb{E}(Z_{f_1}(a))\right)^2\sum_{i=0}^{n^\delta-1}\mathbb{E}\left[\left(\widetilde{N}_n^{(k)}(i,1)\right)^2\right] \leq Ce^{C'a}\left((k+1)^2n^{-\delta}+(k+1)\right) \tag{6.6.8}$$

where $C, C'$ do not depend on $k, a, n, \delta$. Here we use Corollary 6.6.3 in the second inequality. Using (6.6.7) and (6.6.8) to bound the variance term in the right of (6.6.6) completes the proof. ■

The next Lemma provides tight bounds on expectations of descendants of good vertices counted according to $\phi$. Recall $\mu_{f_1}^{(k)}$ denotes the mean measure of a vertex which had degree $k$ at the change point.

**Lemma 6.6.6.** *For any $k\geq 0$,*

$$\varepsilon_n := \left|\mathbb{E}\left[\widetilde{Z}_n^{(k),\phi}\right] - D_n(k,0)\sum_{i=0}^{n^\delta-1} m_{f_1}^{\phi}\left(a-\frac{(i+1)a}{n^\delta}\right)\mu_{f_1}^{(k)}\left[\frac{ia}{n^\delta},\frac{(i+1)a}{n^\delta}\right]\right| \leq Ce^{C'a}\frac{(k+1)D_n(k,0)}{n^\delta}.$$

*Proof.* First note,

$$\mathbb{E}\left[\widetilde{Z}_n^{(k),\phi}\right] = \sum_{i=0}^{n^\delta-1}\sum_{j=1}^{D_n(k,0)}\mathbb{E}\left[\sum_{x\in\widetilde{\mathcal{V}}_n^{(k)}(i,j)} Z_n^{(k),\phi}(i,j,x)\right] = \sum_{i=0}^{n^\delta-1} D_n(k,0)\,\mathbb{E}\left[\mathbb{E}\left[\sum_{x\in\widetilde{\mathcal{V}}_n^{(k)}(i,1)} Z_n^{(k),\phi}(i,1,x)\Big|\mathcal{G}_n\right]\right]$$
$$= D_n(k,0)\sum_{i=0}^{n^\delta-1} m_{f_1}^{\phi}\left(a-\frac{(i+1)a}{n^\delta}\right)\mathbb{E}\left[\widetilde{N}_n^{(k)}(i,1)\right].$$

Here the third equality follows from $\widetilde{\mathcal{V}}_n^{(k)}(i,1)$ is $\mathcal{G}_n$ measurable and for fixed $i$, and for each $x\in\widetilde{\mathcal{V}}_n^{(k)}(i,1)$, conditional on $\mathcal{G}_n$, $Z_n^{(k),\phi}(i,j,x)\stackrel{d}{=}Z_{f_1}^{\phi}\left(a-\frac{(i+1)a}{n^\delta}\right)$. Applying equation (6.6.1), the error term $\varepsilon_n$

in the statement of the Lemma can be bounded as,

$$\varepsilon_n \le 2cD_n(k,0)m_{f_1}(a)\sum_{i=0}^{n^\delta-1}\mathbb{E}\left[N_n^{(k)}(i,1)-\widetilde{N}_n^{(k)}(i,1)\right]. \tag{6.6.9}$$

Next using that the total number of descendants of bad vertices in an interval bounds the number of bad vertices in this interval since each bad vertex has at least one child, we get using Lemma 6.6.4,

$$0 \le \mathbb{E}\left[N_n^{(k)}(i,1)-\widetilde{N}_n^{(k)}(i,1)\right] = \mathbb{E}[B_n^{(k)}(i,1)] \le \mathbb{E}[R_n^{(k)}(i,j)] \le C_1 e^{C_2 a}\frac{(k+1)}{n^{2\delta}}.$$

Using this and (6.6.1) in (6.6.9) completes the proof.

■

**Lemma 6.6.7.** *There exist a constant $a_0 < \infty$ independent of $n,\delta$ such that for any $k \ge 0$, whenever $a \le \frac{\delta}{a_0}\log n$,*

$$\mathbb{E}\left[Z_n^{(k),\phi}-\widetilde{Z}_n^{(k),\phi}\right] \le Ce^{C'a}n^{-\delta}(k+1)D_n(k,0)$$

*Proof.*

$$\mathbb{E}\left[Z_n^{(k),\phi}-\widetilde{Z}_n^{(k),\phi}\right] \le \mathbb{E}\left[\sum_{j=1}^{D_n(k,0)}\sum_{i=0}^{n^\delta-1}\sum_{x\in\mathcal{V}_n^{(k)}(i,j)}Z^{(k),\phi}(i,j,x)\mathbb{1}\{B_x\}\right] = D_n(k,0)\sum_{i=0}^{n^\delta-1}\mathbb{E}\left[\sum_{x\in\mathcal{V}_n^{(k)}(i,1)}Z^{(k),\phi}(i,1,x)\mathbb{1}\{B_x\}\right], \tag{6.6.10}$$

where $B_x$ is the event that a vertex is bad namely has one or more descendants in the interval that it was born. Now note that for a fixed $i$, conditional on the number of births $N_n^{(k)}(i,1)$, we have

$$\sum_{x\in\mathcal{V}_n^{(k)}(i,1)}Z^{(k),\phi}(i,1,x)\mathbb{1}\{B_x\} \preceq_{\mathrm{st}} \sum_{l=1}^{N_n^{(k)}(i,1)}2c|\mathrm{PA}^{(l)}[0,a]|\mathbb{1}\left\{\tilde{B}_l\right\}, \tag{6.6.11}$$

where $\left\{\mathrm{PA}^{(l)} : l \ge 1\right\}$ is a collection of PA branching processes with parameters $\nu = C$ and $\kappa = 0$ (independent of $N_n^{(k)}(i,1)$) and

$$\tilde{B}_l := \left\{\left|\mathrm{PA}^{(l)}\left[0,\frac{a}{n^\delta}\right]\right| \ge 2\right\},$$

namely the root of $\mathrm{PA}^{(l)}$ has at least one child by time $a/n^\delta$. Using this in (6.6.10) implies,

$$\mathbb{E}\left[Z_n^{(k),\phi} - \widetilde{Z}_n^{(k),\phi}\right] \le 2cD_n(k,0)\sum_{i=1}^{n^\delta-1}\mathbb{E}(N_n^{(k)}(i,1))\,\mathbb{E}(|\,\mathrm{PA}^{(1)}[0,a]|\mathbb{1}\left\{\tilde{B}_1\right\}). \tag{6.6.12}$$

Conditioning on the number of births $Y(a/n^\delta)$ of the root of $\mathrm{PA}^{(1)}$ in $[0,a/n^\delta]$ and using the Markov property,

$$\mathbb{E}(|\,\mathrm{PA}^{(1)}[0,a]|\mathbb{1}\left\{\tilde{B}_1\right\}) \le \sum_{j=1}^{\infty}\mathbb{P}\left(Y\left(\frac{a}{n^\delta}\right)=j\right)\mathbb{E}(\mathrm{PA}^{(1),j}[0,a]),$$

where $\mathrm{PA}^{(1),j}$ is a modified PA process with $\nu = C, \kappa = 0$ with the modification that the offspring distribution of the root of $\mathrm{PA}^{(1),j}$ is constructed using attachment function $f(i) := C(j+i+1)$ for $i \ge 0$. Comparing rates, it is easy to see that for each $j \ge 1$, $\mathrm{PA}^{(1),j}[0,a] \preceq_{\mathrm{st}} U_j(a)$, where $U_j(a)$ is constructed by first running a PA processes $\mathrm{PA}_{\nu,\kappa}$ with $\nu = C$ and $\kappa = Cj$ and then setting $U_j(a) = |\mathrm{PA}_{\nu,\kappa}[0,a]|$. By Lemma 6.5.4 for $Y(a/n^\delta)$ and Proposition 6.5.7 for $\mathbb{E}(U_j(a))$, we get $a_0 > 0$ such that whenever $a \le \frac{\delta}{a_0}\log n$,

$$\mathbb{E}(|\,\mathrm{PA}^{(1)}[0,a]|\mathbb{1}\left\{\tilde{B}_1\right\}) \le \sum_{j=1}^{\infty}\left(\frac{Ca}{n^\delta}\right)^j e^{a(2C+Cj)} \le Ce^{C'a}n^{-\delta} \tag{6.6.13}$$

where $C, C'$ do not depend on $k, a, n, \delta$. In (6.6.12), using this bound and using Corollary 6.6.3 for $\mathbb{E}(N_n^{(k)}(i,1))$ completes the proof.

■

**Lemma 6.6.8.** *For any $k \ge 0$, whenever $a \le \frac{\delta}{a_0}\log n$,*

$$\varpi_n := \mathbb{E}\left|Z_n^\phi - \sum_{k=0}^{\infty}D_n(k,0)\sum_{i=0}^{n^\delta-1}m_{f_1}^\phi\left(a-\frac{(i+1)a}{n}\right)\mu_{f_1}^{(k)}\left[\frac{ia}{n^\delta},\frac{(i+1)a}{n^\delta}\right]\right|$$
$$\le Ce^{C'a}\left(n^{1-\delta}+\sqrt{n}+n^{-\delta/2}\left(\sum_{k=1}^{\infty}(k+1)^2D_n(k,0)\right)^{1/2}\right).$$

*Proof.* The term above can be written as $\varpi_n := \varpi_n^{(1)} + \varpi_n^{(2)} + \varpi_n^{(3)}$ where $\varpi_n^{(1)} := Z_n^\phi - \widetilde{Z}_n^\phi$, $\varpi_n^{(2)} := \widetilde{Z}_n^\phi - \mathbb{E}(\widetilde{Z}_n^\phi)$ and

$$\varpi_n^{(3)} := \mathbb{E}(\widetilde{Z}_n^\phi) - \sum_{k=0}^{\infty}D_n(k,0)\sum_{i=0}^{n^\delta-1}m_{f_1}^\phi\left(a-\frac{(i+1)a}{n}\right)\mu_{f_1}^{(k)}\left[\frac{ia}{n^\delta},\frac{(i+1)a}{n^\delta}\right].$$

Now fix $\varepsilon > 0$. Using Lemma 6.6.7 we get,

$$\mathbb{E}(|\varpi_n^{(1)}|) \le \frac{Ce^{C'a}}{n^\delta} \sum_{k=0}^{\infty} (k+1) D_n(k,0) \le 2\gamma C e^{C'a} n^{1-\delta}, \tag{6.6.14}$$

since $\sum_{k=1}^{\infty}(k+1)D_n(k,0) = 2\gamma n - 1$ for tree $\mathcal{T}_{n\gamma}$. Next using Lemma 6.6.5 and Jensen's inequality,

$$\mathbb{E}\left(|\varpi_n^{(2)}|\right) \le Ce^{C'a}\left(\sum_{k=1}^{\infty}\left((k+1)^2 n^{-\delta} + (k+1)\right) D_n(k,0)\right)^{1/2} \le Ce^{C'a}\left(n^{-\delta/2}\left(\sum_{k=1}^{\infty}(k+1)^2 D_n(k,0)\right)^{1/2} + \sqrt{n}\right). \tag{6.6.15}$$

Finally using Lemma 6.6.6 gives,

$$|\varpi_n^{(3)}| \le Ce^{C'a} \sum_{k=0}^{\infty} \frac{(k+1)D_n(k,0)}{n^\delta} \le Ce^{C'a} n^{1-\delta}. \tag{6.6.16}$$

Combining (6.6.14), (6.6.15) and (6.6.16) completes the proof.

■

The next lemma establishes Lipschitz continuity of $m_{f_1}^{\phi}(t)$ in $t$ for any $\phi \in \mathscr{C}$.

**Lemma 6.6.9.** *For any $k \ge 0$ and any $\eta \in [0,1]$,*

$$\sup_{t\in[0,a]} |m_{f_1}^{\phi}(t+\eta) - m_{f_1}^{\phi}(t)| \le Ce^{C'a}\eta.$$

*Proof.* Let $\tau_1$ be the time of the first birth for the branching process with attachment function $f_1$. For any $t \in [0,a]$ and $\eta \in [0,1]$, using the Markov property at time $\eta$, we obtain

$$\begin{aligned} m_{f_1}^{\phi}(t+\eta) = \mathbb{E}\left[Z_{f_1}^{\phi}(t+\eta)\right] &= \mathbb{E}\left[Z_{f_1}^{\phi}(t+\eta)\mathbb{1}\left(\tau_1 > \eta\right)\right] + \mathbb{E}\left[Z_{f_1}^{\phi}(t+\eta)\mathbb{1}\left(\tau_1 \le \eta\right)\right] \\ &= \mathbb{E}\left[Z_{f_1}^{\phi}(t)\right]\mathbb{E}\left[\mathbb{1}\left(\tau_1 > \eta\right)\right] + \mathbb{E}\left[Z_{f_1}^{\phi}(t+\eta)\mathbb{1}\left(\tau_1 \le \eta\right)\right] \\ &= m_{f_1}^{\phi}(t)(1 - \mathbb{P}\left(\tau_1 \le \eta\right)) + \mathbb{E}\left[Z_{f_1}^{\phi}(t+\eta)\mathbb{1}\left(\tau_1 \le \eta\right)\right]. \end{aligned} \tag{6.6.17}$$

Using the strong Markov property at $\tau_1$, we can write the second term above as $\mathbb{E}\left[Z_{f_1}^{\phi}(t+\eta)\mathbb{1}\left(\tau_1 \le \eta\right)\right] = \mathbb{E}\left[\mathbb{E}\left(Z_{f_1}^{\phi}(t+\eta) \mid \mathscr{F}_{\tau_1}\right)\mathbb{1}\left(\tau_1 \le \eta\right)\right]$, where $\mathscr{F}_{\tau_1}$ denotes the associated stopped sigma field. Note that at time $\tau_1$, there are two vertices, one with out-degree one and the other with

out-degree zero. Thus, conditional on $\mathscr{F}_{\tau_1}$, for $i = 1,2$, if $U_i(t)$ is distributed as the size of the PA process $\mathrm{PA}_{\nu,\kappa_i}$ with $\nu = C$ and $\kappa_i = C(i-1)$ at time $t$ (where $C$ is the same constant appearing in Assumption 6.2.1(ii)), we have

$$\mathbb{E}\left(Z_{f_1}^{\phi}(t+\eta) \mid \mathscr{F}_{\tau_1}\right) \le 2c\,\mathbb{E}\left(Z_{f_1}(t+\eta) \mid \mathscr{F}_{\tau_1}\right) \le 2c\,\mathbb{E}(U_1(a+1) + U_2(a+1)) \le Ce^{C'a}$$

for constants $C, C'$ not depending on $\eta, a, t$, where we used Proposition 6.5.7 to get the last inequality. Using this bound and (6.6.1) in (6.6.17), we obtain

$$\begin{aligned}|m_{f_1}^{\phi}(t+\eta) - m_{f_1}^{\phi}(t)| = \left|-m_{f_1}^{\phi}(t)\,\mathbb{P}\left(\tau_1 \le \eta\right) + Ce^{C'a}\,\mathbb{P}\left(\tau_1 \le \eta\right)\right| &\le 2Ce^{C'a}\,\mathbb{P}\left(\tau_1 \le \eta\right) \\ &\le 2Ce^{C'a}(1 - e^{-f_1(0)\eta}) \le C''e^{C'a}\eta\end{aligned}$$

for a constant $C''$ not depending on $\eta, a, t$, where the last equality comes from the fact that $\tau_1 \sim \mathrm{Exp}(f_1(0))$.

■

Now recall $\lambda_k^{\phi}$ defined at the beginning of this Section.

**Lemma 6.6.10.** *For any $k \ge 0$, whenever $a \le \frac{\delta}{a_0}\log n$,*

$$\mathbb{E}\left|Z_n^{\phi} - \sum_{k=0}^{\infty} D_n(k,0)\lambda_k^{\phi}\right| \le Ce^{C'a}\left(n^{1-\delta} + \sqrt{n} + n^{-\delta/2}\left(\sum_{k=1}^{\infty}(k+1)^2 D_n(k,0)\right)^{1/2}\right).$$

*Proof.* Owing to Lemma 6.6.8, it is enough to show, for a positive constants $C, C'$ not depending on $a, n, \delta$ such that

$$\varpi_n^* := \left|\sum_{k=0}^{\infty} D_n(k,0)\lambda_k^{\phi} - \sum_{k=0}^{\infty} D_n(k,0)\sum_{i=0}^{n^{\delta}-1} m_{f_1}^{\phi}\left(a - \frac{(i+1)a}{n}\right)\mu_{f_1}^{(k)}\left[\frac{ia}{n^{\delta}}, \frac{(i+1)a}{n^{\delta}}\right]\right| \le Ce^{C'a}n^{1-\delta}. \tag{6.6.18}$$

Using Lemma 6.6.9,

$$\begin{aligned}\varpi_n^* &\le \sum_{k=0}^{\infty} D_n(k,0) \int_0^a \sum_{i=0}^{n^\delta-1} \left| m_{f_1}^{\phi}(a-s) - m_{f_1}^{\phi}\left(a - \frac{(i+1)a}{n}\right)\right| \mathbb{1}\left(s \in \left[\frac{ia}{n^\delta}, \frac{(i+1)a}{n^\delta}\right]\right) \mu_{f_1}^{(k)}(ds) \\ &\le Ce^{C'a} n^{-\delta} \sum_{k=0}^{\infty} D_n(k,0) \int_0^a \sum_{i=0}^{n^\delta-1} \mathbb{1}\left(s \in \left[\frac{ia}{n^\delta}, \frac{(i+1)a}{n^\delta}\right]\right) \mu_{f_1}^{(k)}(ds) = Ce^{C'a} n^{-\delta} \sum_{k=0}^{\infty} D_n(k,0) \mu_{f_1}^{(k)}[0,a] \\ &\le (Ce^{C'a})^2 an^{-\delta} \sum_{k=0}^{\infty} (k+1) D_n(k,0) = (Ce^{C'a})^2 an^{-\delta}(2\gamma n - 1)\end{aligned}$$

where the last inequality comes from Lemma 6.6.2 and the last equality uses $\sum_{k=0}^{\infty}(k+1)D_n(k,0) = 2\gamma n - 1$.

■

**Lemma 6.6.11.** *Let $\phi \in \mathscr{F}$ then $n \to \infty$,*

$$\sum_{k=1}^{\infty} \frac{D_n(k,0)}{n} \lambda_k^{\phi}(a) \xrightarrow{a.s.} \gamma \sum_{k=1}^{\infty} p_k^0 \lambda_k^{\phi}(a).$$

*Proof.* Let $\chi$ be the characteristic $\chi(t) = \sum_{k=0}^{\infty} \lambda_k^{\phi}(a) \mathbb{1}\{\xi_{f_0}(t) = k\}$. Note by equation (6.6.1) and Lemma 6.6.2 $\lambda_k^{\phi}(a) \le Ce^{C'a}(k+1)$ thus $\chi \in \mathscr{C}$. Now apply Lemma 6.5.8 (i).

■

**Completing the proof of Theorem 6.6.1:** By letting $\delta \to \infty$ keeping $n \ge 2$ fixed in Lemma 6.6.10 the first claim follows. Lemma 6.6.11 then gives the second claim.

## 6.7 Proofs: Sup-norm convergence of degree distribution for the standard model

### 6.7.1 Proof of Theorems 6.3.6 and 6.3.9

In this section, we will prove a convergence result for the empirical degree distribution post change-point. As before, we start time at the change point, i.e. $t = 0$ represents the time $T_{\gamma n}$. Focus will be on the characteristic $\phi(t) = \mathbb{1}\{\xi_{f_1}(t) = k\}$ for $k \ge 0$ and we will denote the corresponding $Z_{f_1}^{\phi}$ and $m_{f_1}^{\phi}$ by $Z_{f_1}^{(k)}$ and $m_{f_1}^{(k)}$ respectively. $\mathrm{BP}_n(t)$ will denote the branching process at time $t$ (i.e. $t$ units after the change point).

#### 6.7.1.1 Notation

We will use the following notation for fixed $t \geq 0$ in this section.

(i) Recall that $n\gamma$ are the number of vertices born **before** the change point. Let $Z_{AC,n}(t)$ = number of vertices at time $t$ who were born **after** the change point. Let $Z_n(t) = n\gamma + Z_{AC,n}(t)$ be the total number of vertices at time $t$.

(ii) Let $\mathscr{D}_n^{BC}(k,t)$ be the set of vertices with degree $k$ at time $t$ who were born **before** the change point $T_{\gamma n}$. Let $D_n^{BC}(k,t) = |\mathscr{D}_n^{BC}(k,t)|$. Similarly, let $\mathscr{D}_n^{AC}(k,t)$ be the set of vertices with degree $k$ at time $t$ who were born **after** the change point. Let $D_n^{AC}(k,t) = |\mathscr{D}_n^{AC}(k,t)|$. Let $D_n(k,t) = D_n^{BC}(k,t) + D_n^{AC}(k,t)$ be the total number of vertices with degree $k$.

(iii) Let $\lambda_\ell^{AC}(t) = \int_0^t m_{f_1}(t-s)\mu_{f_1}^{(\ell)}(ds)$ and $\lambda_\ell^{AC,(k)}(t) = \int_0^t m_{f_1}^{(k)}(t-s)\mu_{f_1}^{(\ell)}(ds)$. Let $\lambda_\ell(t) = 1 + \lambda_\ell^{AC}(t)$ and $\lambda_\ell^{(k)}(t) = \mathbb{P}\left(\xi_{f_1}^{(\ell)}(t) = k-\ell\right) + \lambda_\ell^{AC,(k)}(t)$.

(iv) Let $q_k(t) := \mathbb{P}\left(\xi_{f_1}^{(k)}(t) > 1\right)$.

The following is the main theorem proven in this Section.

**Theorem 6.7.1.** *For any $k \geq 0$, $a > 0$, $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{t\in[0,a]}\left|D_n(k,t) - n\sum_{\ell=0}^{\infty}\gamma p_\ell^0 \lambda_\ell^{(k)}(t)\right| > \epsilon n\right) \to 0$$

*and*

$$\mathbb{P}\left(\sup_{t\in[0,a]}\left|Z_n(t) - n\sum_{\ell=0}^{\infty}\gamma p_\ell^0 \lambda_\ell(t)\right| > \epsilon n\right) \to 0.$$

Assuming the above result for the time being, we now describe how this (coupled with a technical continuity result, Lemma 6.7.4) is now enough to prove Theorems 6.3.6 and 6.3.9. Recall for $m \geq 1$, $T_m = \inf\{t \geq 0 : |\mathrm{BP}_n(t)| = m\}$.

**Corollary 6.7.2.** *Let $G(t) := \sum_{\ell=0}^{\infty} p_\ell^0 \lambda_\ell^{AC}(t)$. For any $s \in [\gamma, 1]$, let $a_s$ be the unique solution to $G(a_s) = \frac{s-\gamma}{\gamma}$ then $n \to \infty$, $\sup_{t\in[\gamma,s]} \left|T_{\lfloor tn \rfloor} - a_t\right| \xrightarrow{\mathrm{P}} 0$.*

*Proof.* As $f_1$ is a strictly positive function, it is easy to see that $G(t)$ is strictly increasing in $t$ and $G(\gamma) = 0$. By Lemma 6.7.4 proved below, $G$ (and hence $G^{-1}$) is continuous. Moreover since $m_{f_1}(t) \geq 1$, $\lambda_\ell^{AC}(t) \geq \mu_{f_1}^{(\ell)}(t) \uparrow \infty$ and we see $G(t) \to \infty$ as $t \to \infty$. Therefore $G(a_s) = \frac{s-\gamma}{\gamma}$ has a unique solution for $s \in [\gamma, 1]$.

Next fix $s \in [\gamma, 1]$ and let $a_s$ be as above. For any $\eta > 0$, choosing $\epsilon = \frac{G(a_s+\eta)-G(a_s)}{2\gamma}$, the second assertion in Theorem 6.7.1 readily implies $\mathbb{P}(Z_n(a_s+\eta) > sn+1) \to 1$. Similarly, it follows that $\mathbb{P}(Z_n(a_s-\eta) < sn-1) \to 1$. Therefore, $T_{\lfloor sn \rfloor} \xrightarrow{\mathrm{P}} a_s$. From this and Theorem 6.7.1, we conclude that $\frac{1}{n}\sup_{t\in[0,T_{\lfloor sn \rfloor}]} \left|Z_n(t) - \gamma n(1+G(t))\right| \xrightarrow{\mathrm{P}} 0$ which implies

$$\sup_{t\in[\gamma,s]} \left| \frac{t-\gamma}{\gamma} - G(T_{\lfloor tn \rfloor}) \right| \xrightarrow{\mathrm{P}} 0.$$

By continuity of $G^{-1}$, this implies

$$\sup_{t\in[\gamma,s]} \left| G^{-1}\left(\frac{t-\gamma}{\gamma}\right) - T_{\lfloor tn \rfloor} \right| \xrightarrow{\mathrm{P}} 0$$

which proves the corollary. ■

*Proof of Theorem 6.3.6.* Fix $s \in [\gamma, 1]$. It follows from Lemma 6.7.4 and Corollary 6.7.6 proved below that $t \mapsto \Phi_t(\mathbf{p}^0)$ is continuous and hence, from Corollary 6.7.2 for each fixed $k \geq 0$,

$$\sup_{t\in[\gamma,s]} \left| \left(\Phi_{T_{\lfloor tn \rfloor}}(\mathbf{p}^0)\right)_k - \left(\Phi_{a_t}(\mathbf{p}^{\mathbf{0}})\right)_k \right| \xrightarrow{\mathrm{P}} 0. \tag{6.7.1}$$

It is easy to see that

$$\begin{aligned} &\sup_{t\in[\gamma,s]} \left| \frac{D_n(k, T_{\lfloor tn \rfloor})}{tn} - \left(\Phi_{T_{\lfloor tn \rfloor}}(\mathbf{p}^{\mathbf{0}})\right)_k \right| \\ &\qquad \leq \frac{1}{\gamma n} \left( \sup_{t\in[0,T_{sn}]} \left| D_n(k,t) - n\sum_{\ell=0}^{\infty} \gamma p_\ell^0 \lambda_\ell^{(k)}(t) \right| + \sup_{t\in[0,T_{sn}]} \left| Z_n(t) - n\sum_{\ell=0}^{\infty} \gamma p_\ell^0 \lambda_\ell(t) \right| \right) \xrightarrow{\mathrm{P}} 0. \end{aligned} \tag{6.7.2}$$

The theorem follows from (6.7.1) and (6.7.2).

■

*Proof of Theorem 6.3.9.* Follows immediately from Theorem 6.3.6. ■

For the remaining portion of this section $C, C', C'', n_0$ will denote generic positive constants not depending on $n, a, k, \ell, t$ whose values might change from line to line. The rest of the Section is devoted to the proof of Theorem 6.7.1.

**Lemma 6.7.3.**

$$q_k(t) \le C(k+1)t$$

*where $C$ is the constant appearing in Assumption 6.2.1(ii) on $f_1$.*

*Proof.* Let $\tau_1^k$ be the time of the first born to a vertex started with degree $k$. Note $\tau_1^k \sim \text{Exp}(f_1(k))$. Thus

$$\mathbb{P}(\tau_1^k < t) = 1 - e^{-f_1(k)t} \le f_1(k)t \le C(k+1)t$$

where the final inequality comes from Assumption 6.2.1(ii) on $f_1$. ■

**Lemma 6.7.4.** *For any $\ell, k \ge 0$ and $t, t+s \le a$,*

$$|\lambda_\ell(t+s) - \lambda_\ell(t)| \le Ce^{C'a}(\ell+1)s, \;\; |\lambda_\ell^{AC,(k)}(t+s) - \lambda_\ell^{AC,(k)}(t)| \le Ce^{C'a}(\ell+1)s.$$

*Proof.* We will only prove the first inequality. The second one follows similarly.

$$\begin{aligned}|\lambda_\ell(t+s) - \lambda_\ell(t)| &\le \int_0^t \left| m_{f_1}(t+s-x) - m_{f_1}(t-x) \right| \mu_{f_1}^{(\ell)}(dx) + \int_t^{t+s} m_{f_1}(t+s-x)\mu_{f_1}^{(\ell)}(dx) \\ &\le Ce^{C'a}s\mathbb{E}\left[\xi_{f_1}^{(\ell)}[0,t]\right] + Ce^{C'a}m_{f_1}(t+s)\mathbb{E}\left[\xi_{f_1}^{(\ell)}[t,t+s]\right] \le Ce^{2C'a}a(\ell+1)s + Ce^{2C'a}(\ell+1)s\end{aligned}$$

where the second inequality uses Lemma 6.6.9 and the third inequality uses Lemma 6.6.2 and (6.6.1). ■

**Lemma 6.7.5.** *For $k \ge l$ and $t, t+s \le a$,*

$$\left| \mathbb{P}\left(\xi_{f_1}^{(\ell)}(t+s) = k-\ell\right) - \mathbb{P}\left(\xi_{f_1}^{(\ell)}(t) = k-\ell\right) \right| \le Ce^{C'a}(k+1)s.$$

*Proof.* We prove this inequality in two steps. First note

$$\begin{aligned}
\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t+s)=k-\ell\right) &= \sum_{d=0}^{k-\ell}\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=d\right)\mathbb{P}\left(\xi_{f_1}^{(d+\ell)}(s)=k-\ell-d\right)\\
&= \sum_{d=0}^{k-\ell-1}\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=d\right)\mathbb{P}\left(\xi_{f_1}^{(d+\ell)}(s)=k-\ell-d\right)+\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\mathbb{P}\left(\xi_{f_1}^{(k)}(s)=0\right)\\
&\leq \sum_{d=0}^{k-\ell-1}\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=d\right)\mathbb{P}\left(\xi_{f_1}^{(d+\ell)}(s)\geq 1\right)+\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\\
&\leq \sum_{d=0}^{k-\ell-1}\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=d\right)\mathbb{E}\,\xi_{f_1}^{(d+\ell)}(s)+\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\\
&\leq \sum_{d=0}^{k-\ell-1}Ce^{C'a}(d+\ell+1)s\,\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=d\right)+\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\\
&\leq Ce^{C'a}s\left(\mathbb{E}\left(\xi_{f_1}^{(\ell)}(t)\right)+\ell+1\right)+\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\\
&\leq C''e^{2C'a}(\ell+1)s+\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right).
\end{aligned}$$

The first equality comes from the Markov property. The second inequality comes from Markov's inequality. The third and fifth inequalities use Lemma Lemma 6.6.2. We now show the opposite inequality.

$$\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t+s)=k-\ell\right)\geq\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\mathbb{P}\left(\xi_{f_1}^{(k)}(s)=0\right)=\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\left(1-\mathbb{P}\left(\xi_{f_1}^{(k)}(s)\geq 1\right)\right)$$

Thus

$$\begin{aligned}
\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t+s)=k-\ell\right)-\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right) &\geq -\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\mathbb{P}\left(\xi_{f_1}^{(k)}(s)\geq 1\right)\\
&\geq -\mathbb{E}\,\xi_{f_1}^{(k)}(s)\geq -Ce^{C'a}(k+1)s
\end{aligned}$$

where the second inequality comes from Markov's inequality and the last inequality comes from Lemma 6.6.2 ■

An immediate consequence of Lemmas 6.7.4 and 6.7.5 is

**Corollary 6.7.6.** *For any $k,\ell>0$ and $t, t+s<a$,*

$$|\lambda_\ell^{(k)}(t+s)-\lambda_\ell^{(k)}(t)|\leq Ce^{C'a}(k+\ell+2)s.$$

**Corollary 6.7.7.** *For any $k$ and $t, t+s<a$,*

$$\sum_{\ell=0}^{\infty} D_n(\ell,0)|\lambda_\ell^{(k)}(t+s)-\lambda_\ell^{(k)}(t)| \le Ce^{C'a}(k+3)sn.$$

*Proof.* By the above Corollary 6.7.6 (with k fixed)

$$\sum_{\ell=0}^{\infty} D_n(\ell,0)|\lambda_\ell^{(k)}(t)-\lambda_\ell^{(k)}(t+s)| \le Ce^{C'a}s\sum_{\ell=0}^{\infty}(k+\ell+2)D_n(\ell,0) \le Ce^{C'a}(k+3)s\gamma n$$

since $\sum_{\ell=0}^{\infty} D_n(\ell,0)=\gamma n$ and $\sum_{\ell=0}^{\infty}\ell D_n(\ell,0)=\gamma n-1$. ■

For the rest of this section, unless specified otherwise, we always work conditional on $\mathscr{F}_n(0)$ so that expectation operations such as $\mathbb{P}(\cdot)$, $\mathbb{E}(\cdot)$ and $\mathrm{Var}(\cdot)$ in the ensuing results mean $\mathbb{P}(\cdot|\mathscr{F}_n(0))$, $\mathbb{E}(\cdot|\mathscr{F}_n(0))$ and $\mathrm{Var}(\cdot|\mathscr{F}_n(0))$ respectively.

We will use Theorem 6.6.1 crucially in what follows for two significant characteristics. Taking $\phi(t)=\mathbb{1}\{t\ge 0\}$ in Theorem 6.6.1, there exist deterministic positive constants $C, C'<\infty$ independent of $a, n$ such that for every $n\ge 2$,

$$\sup_{t\in[0,a]}\mathbb{E}\left|Z_{AC,n}(t)-\sum_{k=0}^{\infty}D_n(k,0)\lambda_k^{AC}(t)\right| < Ce^{C'a}\sqrt{n}. \tag{6.7.3}$$

Taking any $k\ge 0$ and setting $\phi(t)=\mathbb{1}\left\{\xi_{f_1}(t)=k\right\}$ in Theorem 6.6.1, there exist deterministic positive constants $C, C'<\infty$ independent of $a, n, k$ such that for every $n\ge 2$,

$$\sup_{t\in[0,a]}\mathbb{E}\left|D_n^{AC}(k,t)-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{AC,(k)}(t)\right| < Ce^{C'a}\sqrt{n}. \tag{6.7.4}$$

Take any $\widetilde{\theta}\in(0,1/2)$. Take $\omega\in(0,1)$ such that $\omega>\max\left(1-\widetilde{\theta},\frac{1}{2}+\widetilde{\theta}\right)$.

Now let $\{t_i\}_{i=0}^{n^{\widetilde{\theta}}-1}$ be an equispaced partition of $[0,a]$ of mesh $an^{-\widetilde{\theta}}$.

**Lemma 6.7.8.** *Let $\{t_j\},\widetilde{\theta}$ and $\omega$ be as above. Fix $\epsilon\in(0,1)$ and $k$. Then we have*

$$\sum_{j=0}^{n^{\widetilde{\theta}}-1}\mathbb{P}\left(\sup_{t\in[t_j,t_{j+1}]}|D_n(k,t)-D_n(k,t_j)|>\epsilon n^{\omega}\right)\le\frac{Ce^{C'a}}{\epsilon^2}\frac{1}{n^{\omega-\widetilde{\theta}-\frac{1}{2}}}.$$

*Proof.* Condition on $\mathscr{F}_n(t_j)$. Fix $j$ and consider $t\in[t_j,t_{j+1}]$. We clearly have the following lower bound on $D_n(k,t)$:

$$D_n(k,t)\ge D_n(k,t_j)-Y_1$$

where $Y_1$ is the number of degree $k$ vertices at time $t_j$ which have given birth by time $t_{j+1}$. Note that

$$Y_1 \stackrel{\mathrm{d}}{=} \mathrm{Bin}\left(D_n(k,t_j), q_k(an^{-\widetilde{\theta}})\right).$$

We also have the following upper bound on $D_n(k,t)$:

$$D_n(k,t) \le \left(Z_{AC,n}(t_{j+1}) - Z_{AC,n}(t_j)\right) + Y_2 + D_n(k,t_j) \tag{6.7.5}$$

where $Y_2$ denotes the number of vertices existing at time $t_j$ of degree less than $k$ which have given birth by time $t_{j+1}$. Note that

$$Y_2 \stackrel{\mathrm{d}}{=} \sum_{\ell=0}^{k-1} \mathrm{Bin}\left(D_n(\ell,t_j), q_\ell\left(an^{-\widetilde{\theta}}\right)\right).$$

To see this upper bound, note that the degree $k$ vertices at time $t$ originate from vertices either existing at time $t_j$ or new vertices born in the time interval $[t_j,t]$. The latter is bounded by $Z_{AC,n}(t_{j+1}) - Z_{AC,n}(t_j)$, namely, the total number of new births in the time interval $[t_j,t_{j+1}]$. The former is bounded by the sum of the number of vertices which are of degree $k$ at time $t_j$ and have not given birth by time $t$ (which, in turn, is bounded by $D_n(k,t_j)$) and the number of vertices of lower degree at time $t_j$ which have grown to degree $k$ at time $t$ (which, in turn, is bounded by $Y_2$).

These two bounds give the following

$$|D_n(k,t) - D_n(k,t_j)| \le \left(Z_{AC,n}(t_{j+1}) - Z_{AC,n}(t_j)\right) + Y_1 + Y_2.$$

Note the right hand side does not depend on $t$. We now have for all $0 \le j \le n^{\widetilde{\theta}} - 1$ and $t \in [t_j, t_{j+1}]$.

$$\begin{aligned}
&\sup_{j \le n^{\widetilde{\theta}}-1} \mathbb{P}\left(\sup_{t\in[t_j,t_{j+1}]} |D_n(k,t) - D_n(k,t_j)| > \epsilon n^{\omega}\right) \\
&\quad \le \sup_{j \le n^{\widetilde{\theta}}-1} \left[\mathbb{P}\left(\sum_{\ell=0}^{k} \mathrm{Bin}\left(D_n\left(\ell,t_j\right), q_\ell\left(an^{-\widetilde{\theta}}\right)\right) > \frac{\epsilon}{2} n^{\omega}\right) + \mathbb{P}\left(|Z_{AC,n}(t_{j+1}) - Z_{AC,n}(t_j)| > \frac{\epsilon}{2} n^{\omega}\right)\right] \\
&\quad \le \frac{Ce^{C'a}}{\epsilon^2} \frac{1}{n^{\widetilde{\theta}+\omega-\frac{1}{2}}} + \frac{Ce^{C'a}}{\epsilon} \frac{1}{n^{\omega-\frac{1}{2}}}
\end{aligned}$$

where the second inequality comes from Lemmas 6.7.9 and 6.7.10 which are proved below. The result now follows after taking the sum of these terms.

■

**Lemma 6.7.9.** *Let $\{t_j\}, \widetilde{\theta}$ and $\omega$ be as above and let $\epsilon \in (0,1)$. Then there exist constants $C'', n_0$ such that for all $n \geq n_0$ and all $a \leq C'' \log n$,*

$$\sup_{j \leq n^{\tilde{\theta}}} \mathbb{P}\left( \sum_{\ell=0}^{k} Bin\left(D_n(\ell, t_j), q_\ell\left(an^{-\widetilde{\theta}}\right)\right) > \frac{\epsilon}{2} n^{\omega} \right) \leq \frac{Ce^{C'a}}{\epsilon^2} \frac{1}{n^{\widetilde{\theta}+\omega-\frac{1}{2}}}.$$

*Proof.* Define the event $A_j = \left\{ Z_n(t_j) < \left(\gamma + \frac{\epsilon}{8}\right) n^{\widetilde{\theta}+\omega} \right\}$. Note that as $\sum_{\ell=0}^{\infty} (\ell+1) D_n(\ell, t_j) = 2Z_n(t_j) - 1$, therefore on the event $A_j$,

$$\sum_{\ell=0}^{\infty} (\ell+1) D_n(\ell, t_j) < 2\left(\gamma + \frac{\epsilon}{8}\right) n^{\widetilde{\theta}+\omega}. \tag{6.7.6}$$

Applying Chebyshev's inequality, on the event $A_j$, we have

$$\begin{aligned}
\mathbb{P}\left( \sum_{\ell=0}^{k} \mathrm{Bin}\left(D_n(\ell, t_j), q_\ell\left(an^{-\widetilde{\theta}}\right)\right) > \frac{\epsilon}{2} n^{\omega} \middle| \mathscr{F}_n(t_j) \right) &\leq \frac{4}{\epsilon^2 n^{2\omega}} \sum_{\ell=0}^{k} \mathrm{Var}\left( \mathrm{Bin}\left(D_n(\ell, t_j), q_\ell\left(an^{-\widetilde{\theta}}\right)\right) \middle| \mathscr{F}_n(t_j) \right) \\
&\leq \frac{4}{\epsilon^2 n^{2\omega}} \sum_{\ell=0}^{k} D_n(\ell, t_j) q_\ell\left(an^{-\widetilde{\theta}}\right)\left(1 - q_\ell\left(an^{-\widetilde{\theta}}\right)\right) \\
&\leq \frac{4}{\epsilon^2 n^{2\omega}} \frac{Ca}{n^{\widetilde{\theta}}} \sum_{\ell=0}^{k} D_n(\ell, t_j)(\ell+1) \leq \frac{4}{\epsilon^2 n^{2\omega}} \frac{Ca}{n^{\widetilde{\theta}}} \left[ 2\left(\gamma + \frac{\epsilon}{8}\right) n^{\widetilde{\theta}+\omega} \right] \leq \frac{C'a}{\epsilon^2 n^{\omega}}
\end{aligned} \tag{6.7.7}$$

for $C'$ not depnding on $j$, where the first inequality is from Chebyshev's inequality the third inequality is a consequence of Lemma 6.7.3 and the fourth inequality follows from the definition of $A_j$.

We now have

$$\mathbb{P}\left( \sum_{\ell=0}^{k} \mathrm{Bin}\left(D_n(\ell, t_j), q_\ell\left(an^{-\widetilde{\theta}}\right)\right) > \frac{\epsilon}{2} n^{\omega} \right) \leq \frac{C'a}{\epsilon^2 n^{\omega}} + \mathbb{P}\left( Z_n(t_j) \geq \left(\gamma + \frac{\epsilon}{8}\right) n^{\widetilde{\theta}+\omega} \right). \tag{6.7.8}$$

Now, we control the second term above. By Lemma 6.6.2 (and the fact the integral is over a bounded interval) $\lambda_\ell(a) \leq Ce^{C'a}(\ell+1)$. As $\widetilde{\theta} + \omega > 1$, we can clearly choose $C'', n_0$ such that for all $n \geq n_0$ and all $a \leq C'' \log n$, $\frac{\epsilon}{16} n^{\widetilde{\theta}+\omega} > (1+\gamma) Ce^{C'a} n$. For such $n, a$,

$$\sum_{\ell=0}^{\infty} D_n(\ell, 0) \lambda_\ell(t_j) \leq Ce^{C'a} \sum_{\ell=0}^{\infty} (\ell+1) D_n(\ell, 0) = Ce^{C'a}(2\gamma n - 1) < \frac{\epsilon}{16} n^{\widetilde{\theta}+\omega}.$$

Consequently,

$$\begin{aligned}\mathbb{P}\left(Z_n(t_j) \geq \left(\gamma + \frac{\epsilon}{8}\right) n^{\widetilde{\theta}+\omega}\right) &\leq \mathbb{P}\left(Z_n(t_j) - \gamma n \geq \left(\gamma + \frac{\epsilon}{8}\right) n^{\widetilde{\theta}+\omega} - \gamma n\right) \leq \mathbb{P}\left(Z_n(t_j) - \gamma n > \frac{\epsilon}{8} n^{\widetilde{\theta}+\omega}\right) \\ &= \mathbb{P}\left(Z_{AC,n}(t_j) > \frac{\epsilon}{8} n^{\widetilde{\theta}+\omega}\right) \leq \mathbb{P}\left(\left|Z_{AC,n}(t_j) - \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell(t_j)\right| > \frac{\epsilon}{16} n^{\widetilde{\theta}+\omega}\right) \\ &\leq \frac{16}{\epsilon}\frac{1}{n^{\widetilde{\theta}+\omega}}\mathbb{E}\left|Z_{AC,n}(t_j) - \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell(t_j)\right| \leq \frac{16}{\epsilon} C e^{C'a} \frac{1}{n^{\widetilde{\theta}+\omega-\frac{1}{2}}} \quad (6.7.9)\end{aligned}$$

for $C, C'$ not depending on $j$, where the last inequality comes from (6.7.3). (6.7.7) and (6.7.9) and the fact that $\widetilde{\theta} < 1/2$. The result now follows. ■

**Lemma 6.7.10.** *Let $\{t_j\}, \widetilde{\theta}$ and $\omega$ be as above and let $\epsilon > 0$. Then*

$$\sup_{j \leq n^{\widetilde{\theta}}-1} \mathbb{P}\left(\left|Z_{AC,n}(t_{j+1}) - Z_{AC,n}(t_j)\right| > \frac{\epsilon}{2} n^{\omega}\right) \leq \frac{C e^{C'a}}{\epsilon} \frac{1}{n^{\omega-\frac{1}{2}}}.$$

*Proof.* Applying the triangle inequality,

$$\begin{aligned}\left|Z_{AC,n}(t_{j+1}) - Z_{AC,n}(t_j)\right| \leq &\left|Z_{AC,n}(t_{j+1}) - \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell(t_{j+1})\right| + \left|Z_{AC,n}(t_j) - \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell(t_j)\right| \\ &+ \sum_{\ell=0}^{\infty} D_n(\ell,0)\left|\lambda_\ell(t_{j+1}) - \lambda_\ell(t_j)\right|.\end{aligned}$$

Note by Lemma 6.7.4 and the fact that $t_{j+1} - t_j = an^{-\widetilde{\theta}}$

$$\sum_{\ell=0}^{\infty} D_n(\ell,0)\left|\lambda_\ell(t_{j+1}) - \lambda_\ell(t_j)\right| \leq C e^{C'a}\frac{a}{n^{\widetilde{\theta}}}\sum_{\ell=0}^{\infty} D_n(\ell,0)(\ell+1) = C e^{C'a}\frac{a}{n^{\widetilde{\theta}}}(2\gamma n - 1) \leq C'' a e^{C'a} n^{1-\widetilde{\theta}}. \quad (6.7.10)$$

From equation (6.7.3) we get

$$\sup_{j \leq n^{\widetilde{\theta}}-1} \mathbb{E}\left|Z_n(t_j) - \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell(t_j)\right| \leq C e^{C'a}\sqrt{n}.$$

Putting this all together, using (6.7.10) along with the fact that $\omega > (1-\widetilde{\theta})$ and applying Markov's inequality we get for sufficiently large $n$

$$\begin{aligned}
\mathbb{P}\left(\left|Z_{AC,n}(t_{j+1}-Z_{AC,n}(t_j)\right|>\frac{\epsilon}{2}n^{\omega}\right) &= \mathbb{P}\left(\left|Z_n(t_{j+1}-Z_n(t_j)\right|>\frac{\epsilon}{2}n^{\omega}\right)\\
&\leq \mathbb{P}\left(\left|Z_n(t_j)-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell(t_j)\right|+\left|Z_n(t_{j+1})-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell(t_{j+1})\right|>\frac{\epsilon}{4}n^{\omega}\right)\\
&\leq \frac{2}{\epsilon}n^{-\omega}\left(\mathbb{E}\left|Z_n(t_j)-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell(t_j)\right|+\mathbb{E}\left|Z_n(t_{j+1})-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell(t_{j+1})\right|\right)\leq\frac{2Ce^{C'a}}{\epsilon n^{\omega-\frac{1}{2}}}
\end{aligned}$$

for $C, C'$ not depending on $j$, which proves the lemma. ■

**Lemma 6.7.11.** *There exist positive constants $C, C'$ such that for each $k$ and $\epsilon \in (0,1)$,*

$$\mathbb{P}\left(\sup_{t\in[0,a]}\left|D_n(k,t)-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{(k)}(t)\right|>\epsilon(k+1)n^{\omega}\right)\leq\frac{Ce^{C'a}}{\epsilon^2}\frac{1}{n^{\omega-\widetilde{\theta}-\frac{1}{2}}}$$

*and moreover,*

$$\mathbb{P}\left(\sup_{t\in[0,a]}\left|Z_n(t)-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell(t)\right|>\epsilon n^{\omega}\right)\leq\frac{Ce^{C'a}}{\epsilon^2}\frac{1}{n^{\omega-\widetilde{\theta}-\frac{1}{2}}}.$$

*Proof.* Fix $k$ and $\epsilon\in(0,1)$. Note that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,a]}\left|D_n(k,t)-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{(k)}(t)\right|>\epsilon n^{\omega}\right)\\
&\qquad\leq\sum_{j=0}^{n^{\widetilde{\theta}}-1}\mathbb{P}\left(\sup_{t\in[t_j,t_{j+1}]}\left|D_n(k,t)-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{(k)}(t)\right|>\epsilon n^{\omega}\right)\\
&\leq\sum_{j=0}^{n^{\widetilde{\theta}}-1}\left[\mathbb{P}\left(\sup_{t\in[t_j,t_{j+1}]}\left|D_n(k,t)-D_n(k,t_j)\right|>\frac{\epsilon}{3}n^{\omega}\right)+\mathbb{P}\left(\left|D_n(k,t_j)-\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{(k)}(t_j)\right|>\frac{\epsilon}{3}n^{\omega}\right)\right.\\
&\qquad\left.+\mathbb{P}\left(\sup_{t\in[t_j,t_{j+1}]}\sum_{\ell=0}^{\infty}D_n(\ell,0)\left|\lambda_\ell^{(k)}(t)-\lambda_\ell^{(k)}(t_j)\right|>\frac{\epsilon}{3}n^{\omega}\right)\right]. \qquad (6.7.11)
\end{aligned}$$

By Lemma 6.7.8,

$$\sum_{j=0}^{n^{\widetilde{\theta}}-1}\mathbb{P}\left(\sup_{t\in[t_j,t_{j+1}]}\left|D_n(k,t)-D_n(k,t_j)\right|>\frac{\epsilon}{3}n^{\omega}\right)\leq\frac{Ce^{C'a}}{\epsilon^2}\frac{1}{n^{\omega-\widetilde{\theta}-\frac{1}{2}}}. \qquad (6.7.12)$$

By Corollary 6.7.7,

$$\sup_{j\leq n^{\widetilde{\theta}}-1}\sup_{t\in[t_j,t_{j+1}]}\sum_{\ell=0}^{\infty}D_n(\ell,0)\left|\lambda_\ell^{(k)}(t)-\lambda_\ell^{(k)}(t_j)\right|\leq Ce^{C'a}(k+\gamma+2)n^{1-\widetilde{\theta}}$$

and hence, as $\omega > 1-\widetilde{\theta}$, there exists $n_0$ not depending on $k$ such that for all $n \geq n_0$,

$$\sum_{j=0}^{n^{\tilde{\theta}}-1} \mathbb{P}\left( \sup_{t\in[t_j,t_{j+1}]} \sum_{\ell=0}^{\infty} D_n(\ell,0) \left| \lambda_\ell^{(k)}(t) - \lambda_\ell^{(k)}(t_j) \right| > \frac{\epsilon(k+1)}{3} n^{\omega} \right) = 0. \tag{6.7.13}$$

Finally we control the second term appearing in the sum (6.7.11). It is sufficient to show

$$\sup_{j\leq n^{\tilde{\theta}}} \mathbb{P}\left( \left| D_n(k,t_j) - \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{(k)}(t_j) \right| > \frac{\epsilon}{3} n^{\omega} \right) \leq \frac{Ce^{C'a}}{\epsilon^2} \frac{1}{n^{\omega-\frac{1}{2}}}. \tag{6.7.14}$$

By the triangle inequality and definitions of $D_n(k,t)$, and $\lambda_\ell^{(k)}(t)$, we see that for each fixed $j,k$,

$$\begin{aligned} \left| D_n(k,t_j) - \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{(k)}(t_j) \right| \leq & \left| D_n^{BC}(k,t_j) - \sum_{\ell=0}^{k} D_n(\ell,0)\,\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t_j) = k-\ell\right) \right| \\ & + \left| D_n^{AC}(k,t_j) - \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{AC,(k)}(t_j) \right|. \end{aligned} \tag{6.7.15}$$

By (6.7.4) and Markov's inequality,

$$\sup_{j\leq n^{\tilde{\theta}}} \mathbb{P}\left( \left| D_n^{AC}(k,t_j) - \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{AC,(k)}(t_j) \right| > \frac{\epsilon}{6} n^{\omega} \right) \leq \frac{6Ce^{C'a}}{\epsilon} \frac{1}{n^{\omega-\frac{1}{2}}}. \tag{6.7.16}$$

We now control the first term appearing in the bound in equation (6.7.15) by showing

$$\sup_{t\in[0,a]} \mathbb{E}\left[ \left( D_n^{BC}(k,t) - \sum_{\ell=0}^{k} D_n(\ell,0)\,\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t) = k-\ell\right) \right)^2 \right] \leq Cn. \tag{6.7.17}$$

Fix $k$ and $t \in [0,a]$. Define a collection of mutually independent random variables $\left\{ \xi_{f_1,m}^{(\ell)}(t) \mid 1 \leq m \leq D_n(\ell,0), 0 \leq \ell \leq k \right\}$ where $\xi_{f_1,m}^{(\ell)}(t) \sim \xi_{f_1}^{(\ell)}(t)$. Note that

$$D_n^{BC}(k,t) \stackrel{d}{=} \sum_{\ell=0}^{k} \sum_{m=1}^{D_n(\ell,0)} \mathbb{1}\left(\xi_{f_1,m}^{(\ell)}(t) = k-\ell\right),$$

i.e. a vertex that was born before the change point and was of degree $\ell$ at the change point has to add $k-\ell$ new births to reach degree $k$ at time $t$.

Therefore,

$$\mathbb{E}\left[\left(D_n^{BC}(k,t)-\sum_{\ell=0}^{k}D_n(\ell,0)\,\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\right)^2\right]$$
$$=\mathbb{E}\left[\left(\sum_{\ell=0}^{k}\sum_{m=1}^{D_n(\ell,0)}\mathbb{1}\left(\xi_{f_1,m}^{(\ell)}(t)=k-\ell\right)-\sum_{\ell=0}^{k}D_n(\ell,0)\,\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\right)^2\right]$$
$$=\mathbb{E}\left[\left\{\sum_{\ell=0}^{k}\sum_{m=1}^{D_n(\ell,0)}\left(\mathbb{1}\left(\xi_{f_1,m}^{(\ell)}(t)=k-\ell\right)-\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\right)\right\}^2\right].$$

Note that

$$\sum_{\ell=0}^{k}\sum_{m=1}^{D_n(\ell,0)}\left(\mathbb{1}\left(\xi_{f_1,m}^{(\ell)}(t)=k-\ell\right)-\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t)=k-\ell\right)\right)\stackrel{d}{=}\sum_{\ell=0}^{k}\sum_{m=1}^{D_n(\ell,0)}Y_{\ell,m}$$

Where the random variables $\left\{Y_{\ell,m}\mid 1\le m\le D_n(\ell,0),0\le\ell\le k\right\}$ are mutually independent, supported on $[-1,1]$ and $\mathbb{E}\,Y_{\ell,m}=0$. Thus,

$$\mathbb{E}\left[\left(\sum_{\ell=0}^{k}\sum_{m=1}^{D_n(\ell,0)}Y_{\ell,m}\right)^2\right]=\sum_{\ell=0}^{k}\sum_{m=1}^{D_n(\ell,0)}\mathbb{E}\left[Y_{\ell,m}^2\right]\le C\sum_{\ell=0}^{k}D_n(\ell,0)=C\gamma n$$

which proves (6.7.17). Using (6.7.17) and Chebychev's inequality, we get

$$\sup_{j\le n^{\tilde{\theta}}}\mathbb{P}\left(\left|D_n^{BC}(k,t_j)-\sum_{\ell=0}^{k}D_n(\ell,0)\,\mathbb{P}\left(\xi_{f_1}^{(\ell)}(t_j)=k-\ell\right)\right|>\frac{\epsilon}{6}n^{\omega}\right)\le\frac{C}{\epsilon^2 n^{2\omega-1}}.\tag{6.7.18}$$

Using (6.7.16) and (6.7.18) in (6.7.15), we obtain (6.7.14). The first assertion in the lemma follows by using (6.7.12), (6.7.13) and (6.7.14) in (6.7.11). The second assertion follows similarly upon noting that $Z_{AC,n}(t)$ is increasing in $t$ and using (6.7.3), Lemma 6.7.10 and the first bound in Lemma 6.7.4.

■

Now, we proceed towards removing the conditioning on $\mathscr{F}_n(0)$ to complete the proof of Theorem 6.7.1. We need the following Corollary to Lemma 6.6.11.

**Corollary 6.7.12.** *Fix $k\ge 0$, $\epsilon>0$ and let $s_1,\ldots,s_m\in[0,a]$ be $m$ fixed time points. Then, almost surely, there exists $n_0\ge 1$ such that that for all $n\ge n_0$,*

$$\sup_{1\le j\le m}\left|\frac{1}{n}\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{(k)}(s_j)-\gamma\sum_{\ell=0}^{\infty}p_\ell^0\lambda_\ell^{(k)}(s_j)\right|\le\epsilon.$$

*Moreover,*

$$\sup_{1\le j\le m}\left|\frac{1}{n}\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell(s_j)-\gamma\sum_{\ell=0}^{\infty}p_\ell^0\lambda_\ell(s_j)\right|\le\epsilon.$$

*Proof.* Follows from Lemma 6.6.11 and the union bound. ■

**Lemma 6.7.13.** *Let* $\{p_k(f):k\ge 0\}$ *as in* (6.3.1) *be the asymptotic degree distribution using attachment function* $f$ *satisfying Assumption 6.2.1. Then* $\sum_{k=0}^{\infty}kp_k(f)=1$.

*Proof.* Recall that $p_k(f)=t_{k-1}-t_k$ where $t_k:=\prod_{i=0}^{k}\frac{f(i)}{\lambda^*+f(i)}$ and $\lambda^*$ is the Malthusian parameter for the corresponding preferential attachment branching process. Therefore, $\sum_{k=1}^{\infty}kp_k(f)=\sum_{k=0}^{n}k(t_{k-1}-t_k)=\sum_{k=0}^{\infty}t_k$. By the definition of $\lambda^*$ and $t_k$ we see $\sum_{k=1}^{\infty}t_k=1$, proving the lemma.

■

**Lemma 6.7.14.** *For any* $k\ge 0$,

$$\sup_{t\in[0,a]}\left|\frac{1}{n}\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{(k)}(t)-\gamma\sum_{\ell=0}^{\infty}p_\ell^0\lambda_\ell^{(k)}(t)\right|\xrightarrow{a.s.}0.$$

*Moreover,*

$$\sup_{t\in[0,a]}\left|\frac{1}{n}\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell(t)-\gamma\sum_{\ell=0}^{\infty}p_\ell^0\lambda_\ell(t)\right|\xrightarrow{a.s.}0.$$

*Proof.* Fix $\epsilon>0$. Let $0=s_1<s_2<\cdots<s_m=a$ be a partition such that $|s_{j+1}-s_j|\le\epsilon$.

By Corollary 6.7.7,

$$\sup_{1\le j\le m}\sup_{t\in[s_j,s_{j+1}]}\left|\frac{1}{n}\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{(k)}(t)-\frac{1}{n}\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{(k)}(s_j)\right|\le Ce^{C'a}(k+3)\epsilon.$$

Similarly, using Corollary 6.7.6,

$$\begin{aligned}\sup_{1\le j\le m-1}\sup_{t\in[s_j,s_{j+1}]}\left|\gamma\sum_{\ell=0}^{\infty}p_\ell^0\lambda_\ell^{(k)}(t)-\gamma\sum_{\ell=0}^{\infty}p_\ell^0\lambda_\ell^{(k)}(s_j)\right|&\le\sup_{1\le j\le k-1}\sup_{[s_j,s_{j+1}]}\gamma\sum_{\ell=0}^{\infty}p_\ell^0\left|\lambda_\ell^{(k)}(t)-\lambda_\ell^{(k)}(s_j)\right|\\&\le Ce^{C'a}\epsilon\gamma\sum_{\ell=0}^{\infty}p_\ell^0(k+\ell+2)=Ce^{C'a}\gamma(k+3)\epsilon.\end{aligned}$$

By Corollary 6.7.12, almost surely, there exists $n_0\ge 1$ such that that for all $n\ge n_0$,

$$\sup_{1\le j\le m}\left|\frac{1}{n}\sum_{\ell=0}^{\infty}D_n(\ell,0)\lambda_\ell^{(k)}(s_j)-\gamma\sum_{\ell=0}^{\infty}p_\ell^0\lambda_\ell^{(k)}(s_j)\right|\le\epsilon.$$

From the above, we now have that for $n \geq n_0$,

$$\begin{aligned}
&\sup_{t\in[0,a]}\left|\frac{1}{n}\sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{(k)}(t) - \gamma\sum_{\ell=0}^{\infty} p_\ell^0\lambda_\ell^{(k)}(t)\right| \\
&\qquad\leq \sup_{1\leq j\leq m-1}\sup_{t\in[s_j,s_{j+1}]}\left|\frac{1}{n}\sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{(k)}(t) - \frac{1}{n}\sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{(k)}(s_j)\right| \\
&+ \sup_{1\leq j\leq m-1}\sup_{t\in[s_j,s_{j+1}]}\left|\gamma\sum_{\ell=0}^{\infty} p_\ell^0\lambda_\ell^{(k)}(t) - \gamma\sum_{\ell=0}^{\infty} p_\ell^0\lambda_\ell^{(k)}(s_j)\right| + \sup_{1\leq j\leq m}\left|\frac{1}{n}\sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{(k)}(s_j) - \gamma\sum_{\ell=0}^{\infty} p_\ell^0\lambda_\ell^{(k)}(s_j)\right| \\
&\qquad\leq Ce^{C'a}(k+3)\epsilon
\end{aligned}$$

which proves the first assertion of the lemma. The second assertion follows similarly using Corollary 6.7.12 and the first bound in Lemma 6.7.4.

■

*Proof of Theorem 6.7.1.* The theorem follows from Lemmas 6.7.11 and 6.7.14.

■

### 6.7.2 Proof of Corollary 6.3.11:

The essential message of this Corollary 6.3.11 is that the tail of the distribution prescribed by the initializer function always wins. Recall that the limit random variable $D_{\boldsymbol{\theta}}$ is a mixture of the distributions of $X_{\mathsf{BC}}$ and $X_{\mathsf{AC}}$.

**Lemma 6.7.15.** *The random variable $X_{\mathsf{AC}}$ always has an exponential tail.*

**Proof:** By construction, note that $X_{\mathsf{AC}} \preceq_{\text{st}} \xi_{f_1}[0,\alpha]$. Further our assumption on the attachment functions implies that there exists $\kappa > 0$ such that $\max(f_0(i), f_1(i)) \leq \kappa(i+1)$ for all $i$. In particular $\xi_{f_1}[0,\alpha] \preceq_{\text{st}} Y_\kappa[0,\alpha]$ where $Y_\kappa(\cdot)$ is a rate $\kappa$ Yule process as in Definition 6.5.3. Using Lemma 6.5.4 now completes the proof.

■

Thus is is enough to consider $X_{\mathsf{BC}}$ and show that this random variable has the same tail behavior as the random variable $D \sim \{p_k^0 : k \geq 1\}$. Once again by construction,

$$X_{\mathsf{BC}} \preceq_{\text{st}} D + \sum_{i=1}^{D} Y_{\kappa,i}[0,\alpha],$$

where $\{Y_{\kappa,i}(\cdot) : i \geq 1\}$ is an infinite collection of independent Yule processes (independent of $D$). Let $\mu := \mathbb{E}(Y_{\kappa,i}[0,\alpha])$. Note $\mu > 1$. Now note that for $x \geq 1$,

$$\begin{aligned}\mathbb{P}(X_{\mathsf{BC}} > x) &\leq \sum_{j=1}^{x/2\mu} \mathbb{P}(D=j)\,\mathbb{P}\left(\sum_{i=1}^{j} Y_{\kappa,i}[0,\alpha] > x-j\right) + \mathbb{P}(D > x/2\mu) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{x/2\mu} Y_{\kappa,i}[0,\alpha] > x\left(1-\frac{1}{2\mu}\right)\right) + \mathbb{P}(D > x/2\mu). \end{aligned} \tag{6.7.19}$$

Standard large deviation bounds for the probability measure of $Y_{\kappa,i}$ implies that there exists constants $C_1, C_2$ such that for all $x$,

$$\mathbb{P}\left(\sum_{i=1}^{x/2\mu} Y_{\kappa,i}[0,\alpha] > x\left(1-\frac{1}{2\mu}\right)\right) \leq C_1 \exp(-C_2 x).$$

Thus in the setting of Corollary 6.3.11(i), assuming $D$ has exponential tails, one finds using (6.7.19) that there exist finite constants $C_1', C_2'$ such that

$$\mathbb{P}(X_{\mathsf{BC}} > x) \leq C_1' \exp(-C_2' x).$$

This completes the proof of Corollary 6.3.11(i). A similar argument using the obvious inequality $\mathbb{P}(D > x) \leq \mathbb{P}(X_{\mathsf{BC}} > x)$ verifies Corollary 6.3.11(ii).

■

## 6.8 Proofs: Quick Big bang

### 6.8.1 Proof of Theorem 6.3.15

Recall that in this section, we throughout work under Assumptions 6.2.1, 6.3.1 and 6.3.14 for $f_0, f_1$. For notational convenience, instead of considering the change point at $T_{n^\gamma}$ and evolving the tree till $T_n$, we will consider the problem of the change point being at $T_n$ and evolving the tree till $T_{n^{1+\lambda_1^*\theta}}$ for some $\theta > 0$ (where $\lambda_1^*$ is the Malthusian rate corresponding to $f_1$). For this section, $t = 0$ represents time $T_n$ (the smallest time the change point process has $n$ vertices). It is easy to see that Theorem 6.3.15 is equivalent to Theorem 6.8.13 proved below.

Recall the notation from Section 6.7. From Lemma 6.7.11, for every $k \geq 0$, there exists $\eta_0 > 0$ such that for all $\eta \leq \eta_0$,

$$\frac{1}{n} \sup_{t \in [0, \eta \log n]} \left| D_n(k,t) - \sum_{\ell=0}^{\infty} D_n(\ell,0) \lambda_\ell^{(k)}(t) \right| \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{6.8.1}$$

Similarly, using Lemma 6.7.11, we obtain $\eta_0$ such that for all $\eta \leq \eta_0$,

$$\frac{1}{n} \sup_{t \in [0, \eta \log n]} \left| Z_n(t) - \sum_{\ell=0}^{\infty} D_n(\ell,0) \lambda_\ell(t) \right| \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{6.8.2}$$

(6.8.1) and (6.8.2) immediately imply for any $\eta \leq \eta_0$,

$$\begin{aligned} \frac{1}{n^{1+\eta\lambda_1^*}} D_n(k, \eta \log n) - \frac{1}{n^{1+\eta\lambda_1^*}} \sum_{\ell=0}^{\infty} D_n(\ell,0) \lambda_\ell^{(k)}(\eta \log n) &\xrightarrow{P} 0, \\ \frac{1}{n^{1+\eta\lambda_1^*}} Z_n(\eta \log n) - \frac{1}{n^{1+\eta\lambda_1^*}} \sum_{\ell=0}^{\infty} D_n(\ell,0) \lambda_\ell(\eta \log n) &\xrightarrow{P} 0 \end{aligned} \tag{6.8.3}$$

as $n \to \infty$. Define for each $\ell \geq 0$ and $\beta > 0$,

$$w_\ell(\beta) := \int_0^\infty e^{-\beta s} \mu_{f_1}^{(\ell)}(ds).$$

We will simply write $w_\ell$ for $w_\ell(\lambda_1^*)$. We will need the following technical lemmas. Recall from Assumption 6.2.1 (iii) that there exists $\beta_1 \in (0, \lambda_1^*)$ such that $\hat{\rho}(\beta_1) < \infty$. Recall $C^*$ from Assumption 6.3.1 applied to $f_1$.

**Lemma 6.8.1.** $\beta_1 \geq C^*$.

*Proof.* If $C^* = 0$, there is nothing to prove. So we assume $C^* > 0$. For any $\epsilon \in (0, C^*)$, by Assumption 6.3.1, there exists $j_0 \geq 1$ such that for all $j \geq j_0$, $f_1(j) \geq (C^* - \epsilon) j$. Finiteness of $\hat{\rho}(\beta_1)$ implies that

$$\sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{f_1(i + j_0)}{\beta_1 + f_1(i + j_0)} < \infty. \tag{6.8.4}$$

For any $k \geq 1$, noting that $x \mapsto \frac{x}{\beta_1 + x}$ is a strictly increasing function and, $\log(1+x) \leq x$ for any $x \geq 0$, and $\sum_{j=j_1}^{j_2} \frac{1}{j} \leq \int_{j_1-1}^{j_2} \frac{dx}{x}$ for any $j_2 \geq j_1 \geq 1$,

$$\log\left[\prod_{i=0}^{k-1} \frac{f_1(i+j_0)}{\beta_1 + f_1(i+j_0)}\right] \geq \log\left[\prod_{i=0}^{k-1} \frac{i+j_0}{\frac{\beta_1}{C^*-\epsilon} + i + j_0}\right] = -\sum_{i=0}^{k-1} \log\left[1 + \frac{\beta_1}{(C^*-\epsilon)(i+j_0)}\right]$$
$$\geq -\frac{\beta_1}{C^*-\epsilon} \sum_{i=0}^{k-1} \frac{1}{i+j_0} \geq -\frac{\beta_1}{C^*-\epsilon} \int_{j_0-1}^{j_0+k-1} \frac{dx}{x} = -\frac{\beta_1}{C^*-\epsilon} \log\left(\frac{j_0+k-1}{j_0-1}\right)$$

and thus

$$\prod_{i=0}^{k-1} \frac{f_1(i+j_0)}{\beta_1 + f_1(i+j_0)} \geq \left(\frac{j_0-1}{j_0+k-1}\right)^{\frac{\beta_1}{C^*-\epsilon}}.$$

Thus, (6.8.4) holds only if $\beta_1 > C^* - \epsilon$. As $\epsilon > 0$ is arbitrary, this proves the lemma.

■

**Lemma 6.8.2.** *For any $\beta \in (\beta_1, \lambda_1^*]$, there exists a constant $C(\beta) > 0$ such that $w_\ell(\beta) \leq C(\beta)(\ell+1)$ for all $\ell \geq 0$.*

*Proof.* Fix any $\beta \in (\beta_1, \lambda_1^*]$ and $\ell \geq 0$. Since $\int_0^\infty e^{-\beta s} \mu_{f_1}(ds) = \sum_{k=1}^\infty \prod_{i=0}^{k-1} \frac{f_1(i)}{\beta + f_1(i)}$, the sum on the right hand side is finite. Note that

$$w_\ell(\beta) = \int_0^\infty e^{-\beta s} \mu_{f_1}^{(\ell)}(ds) = \sum_{k=1}^\infty \prod_{i=\ell}^{\ell+k-1} \frac{f_1(i)}{\beta + f_1(i)} = \frac{\sum_{k=1}^\infty \prod_{i=0}^{\ell+k-1} \frac{f_1(i)}{\beta+f_1(i)}}{\prod_{i=0}^{\ell-1} \frac{f_1(i)}{\beta+f_1(i)}} < \infty.$$

Choose and fix $\epsilon > 0$ such that $C^* + 2\epsilon < \beta$ (which is possible by Lemma 6.8.1). By Assumption 6.3.1, there exists $j_0 \geq 1$ such that for all $j \geq j_0$, $f_1(j) \leq (C^* + \epsilon) j$. For any $\ell \geq j_0$, using the facts that $x \mapsto \frac{x}{\beta + x}$ is a strictly increasing function and, $\log(1+x) \geq \frac{x}{1+x}$ for any $x \geq 0$, and $\sum_{j=j_1}^{j_2} \frac{1}{j} \geq \int_{j_1}^{j_2+1} \frac{dx}{x}$ for any $j_2 \geq j_1 \geq 1$, we obtain for any $\ell \geq j_0$,

$$\log\left[\prod_{i=\ell}^{2\ell-1} \frac{f_1(i)}{\beta + f_1(i)}\right] \leq \log\left[\prod_{i=\ell}^{2\ell-1} \frac{i}{\frac{\beta}{C^*+\epsilon} + i}\right] = -\sum_{i=\ell}^{2\ell-1} \log\left[1 + \frac{\beta}{(C^*+\epsilon)i}\right]$$
$$\leq -\sum_{i=\ell}^{2\ell-1} \frac{\frac{\beta}{(C^*+\epsilon)i}}{1 + \frac{\beta}{(C^*+\epsilon)i}} \leq -\frac{\frac{\beta}{C^*+\epsilon}}{1 + \frac{\beta}{(C^*+\epsilon)\ell}} \sum_{i=\ell}^{2\ell-1} \frac{1}{i} \leq -\frac{\frac{\beta}{C^*+\epsilon}}{1 + \frac{\beta}{(C^*+\epsilon)\ell}} \int_\ell^{2\ell} \frac{dx}{x} = -\frac{\frac{\beta}{C^*+\epsilon}}{1 + \frac{\beta}{(C^*+\epsilon)\ell}} \log 2.$$

Take $\ell_1 \geq j_0$ such that $\frac{\frac{\beta}{C^*+\epsilon}}{1+\frac{\beta}{(C^*+\epsilon)\ell_1}} \geq \frac{\beta}{C^*+2\epsilon}$. From the above calculation, for all $\ell \geq \ell_1$, $\prod_{i=\ell}^{2\ell-1} \frac{f_1(i)}{\beta+f_1(i)} \leq 2^{-\frac{\beta}{C^*+2\epsilon}}$. Using this bound iteratively, we obtain for any $j \geq 1$,

$$\prod_{i=\ell}^{2^j\ell-1} \frac{f_1(i)}{\beta+f_1(i)} \leq 2^{-\frac{\beta j}{C^*+2\epsilon}}.$$

Thus, for all $\ell \geq \ell_1$,

$$\begin{aligned} w_\ell(\beta) = \sum_{k=1}^{\infty} \prod_{i=\ell}^{\ell+k-1} \frac{f_1(i)}{\beta+f_1(i)} \leq \ell + \sum_{j=0}^{\infty} \sum_{k=2^j\ell}^{2^{j+1}\ell-1} \prod_{i=\ell}^{\ell+k-1} \frac{f_1(i)}{\beta+f_1(i)} \leq \ell + \sum_{j=0}^{\infty} 2^j\ell \prod_{i=\ell}^{2^j\ell-1} \frac{f_1(i)}{\beta+f_1(i)} \\ = \ell\left[1+\sum_{j=0}^{\infty} 2^{\left(1-\frac{\beta}{C^*+2\epsilon}\right)j}\right] = \left(\frac{2-2^{\left(1-\frac{\beta}{C^*+2\epsilon}\right)}}{1-2^{\left(1-\frac{\beta}{C^*+2\epsilon}\right)}}\right)\ell \end{aligned}$$

where the sum converges as $C^* + 2\epsilon < \beta$. This proves the lemma.

■

Recall the class of characteristics $\mathscr{C}$ defined in (6.3.2).

**Lemma 6.8.3.** *Let $\phi \in \mathscr{C}$ such that $\lim_{t\to\infty} e^{-\lambda_1^* t} m_{f_1}^{\phi}(t) = c_\phi$. For $\ell \geq 0$, define*

$$\lambda_\ell^{\phi}(t) = \lambda_\ell^{\phi}(0) + \int_0^t m_{f_1}^{\phi}(t-s)\mu_{f_1}^{(\ell)}(ds) \tag{6.8.5}$$

*where $\lambda_\ell^{\phi}(0) \in [0,1]$ for each $\ell$. There is a constant $C > 0$ for which the following holds: for any $\epsilon > 0$, there exists $t(\epsilon) > 0$ such that for any $\ell \geq 0$,*

$$\sup_{t \leq t(\epsilon)} \left| e^{-\lambda_1^* t} \lambda_\ell^{\phi}(t) - w_\ell c_\phi \right| \leq C\epsilon(\ell+1).$$

*Proof.* In this proof, $C, C', C''$ will denote generic positive constants not depending on $t, \ell$ whose values might change from line to line. From (6.8.5) and the definition of $w_\ell$, we have for any $t \geq 0$,

$$e^{-\lambda_1^* t}\lambda_\ell^{\phi}(t) - w_\ell c_\phi = \lambda_\ell^{\phi}(0)e^{-\lambda_1^* t} - c_\phi \int_t^{\infty} e^{-\lambda_1^* s}\mu_{f_1}^{(\ell)}(ds) + \int_0^t \left(e^{-\lambda_1^*(t-s)} m_{f_1}^{\phi}(t-s) - c_\phi\right) e^{-\lambda_1^* s}\mu_{f_1}^{(\ell)}(ds). \tag{6.8.6}$$

Choose any $\epsilon > 0$. Take and fix any $\vartheta > 0$ such that $\lambda_1^* - \vartheta > \beta_1$. As $\lim_{t\to\infty} e^{-\lambda_1^* t} m_{f_1}^{\phi}(t) = c_\phi$ and $\sup_{t<\infty} e^{-\lambda_1^* t} m_{f_1}^{\phi}(t) < \infty$ (which holds because the limit as $t \to \infty$ exists and as $\phi \in \mathscr{C}$, therefore for each

$a > 0$, $\sup_{t\in[0,a]} m_{f_1}^{\phi}(t) \le C \sup_{t\in[0,a]} m_{f_1}(t) < \infty$ by virtue of (6.6.1)), there exists $t_0 > 0$ such that for all $t \ge t_0$, $\left|e^{-\lambda_1^* t} m_{f_1}^{\phi}(t) - c_\phi\right| \le \epsilon$ and $e^{-\vartheta t}\left(\sup_{z<\infty} e^{-\lambda_1^* z} m_{f_1}^{\phi}(z) + c_\phi\right) \le \epsilon$. Thus, for any $t \ge 2t_0$,

$$\sup_{s\le t} e^{-\vartheta s}\left|e^{-\lambda_1^*(t-s)} m_{f_1}^{\phi}(t-s) - c_\phi\right| \le \epsilon.$$

Thus, applying Lemma 6.8.2 with $\beta = \lambda_1^* - \vartheta$, we conclude that for any $t \ge 2t_0$,

$$\begin{aligned}\int_0^t \left|e^{-\lambda_1^*(t-s)} m_{f_1}^{\phi}(t-s) - c_\phi\right| e^{-\lambda_1^* s} \mu_{f_1}^{(\ell)}(ds) \\ = \int_0^t e^{-\vartheta s}\left|e^{-\lambda_1^*(t-s)} m_{f_1}^{\phi}(t-s) - c_\phi\right| e^{-(\lambda_1^*-\vartheta)s} \mu_{f_1}^{(\ell)}(ds) \le \epsilon w_\ell(\lambda_1^* - \vartheta) \le C\epsilon(\ell+1).\end{aligned}$$

Moreover, as $\int_0^\infty e^{-(\lambda_1^*-\vartheta)s}\mu_{f_1}^{(\ell)}(ds) \le C(\ell+1)$, for $t \ge 0$,

$$c_\phi \int_t^\infty e^{-\lambda_1^* s}\mu_{f_1}^{(\ell)}(ds) \le C'(\ell+1)e^{-\vartheta t}.$$

Using these in (6.8.6) and recalling $\lambda_\ell(0) \in [0,1]$ for each $\ell$, we obtain for $t \ge 2t_0$,

$$\left|e^{-\lambda_1^* t}\lambda_\ell^{\phi}(t) - w_\ell c_\phi\right| \le e^{-\lambda_1^* t} + C'(\ell+1)e^{-\vartheta t} + C\epsilon(\ell+1).$$

Thus, there exists $t_1 \ge 2t_0$ such that for all $\ell \ge 0$ and all $t \ge t_1$,

$$\left|e^{-\lambda_1^* t}\lambda_\ell^{\phi}(t) - w_\ell c_\phi\right| \le C''\epsilon(\ell+1).$$

■

**Lemma 6.8.4.** *Let $\phi \in \mathscr{C}$ such that $\lim_{t\to\infty} e^{-\lambda_1^* t} m_{f_1}^{\phi}(t) = c_\phi$. For $\ell \ge 0$, let $\lambda_\ell^{\phi}(\cdot)$ be defined as in* (6.8.5). *Fix any $\eta > 0, a \in \mathbb{R}$. Then as $n \to \infty$,*

$$\frac{1}{n^{1+\eta\lambda_1^*}} \sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{\phi}(\eta\log n + a) \xrightarrow{P} c_\phi e^{\lambda_1^* a} \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell.$$

*Proof.* In this proof, $C, C', C''$ will denote generic positive constants not depending on $n, t, \ell$ whose values might change from line to line. Note that

$$\left|\frac{1}{n^{1+\eta\lambda_1^*}}\sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^\phi(\eta\log n+a) - c_\phi e^{\lambda_1^* a}\sum_{\ell=0}^{\infty} p_\ell^0 w_\ell\right|$$
$$\leq \sum_{\ell=0}^{\infty}\frac{D_n(\ell,0)}{n}\left|\frac{\lambda_\ell^\phi(\eta\log n+a)}{n^{\eta\lambda_1^*}} - w_\ell c_\phi e^{\lambda_1^* a}\right| + c_\phi e^{\lambda_1^* a}\left|\sum_{\ell=0}^{\infty}\frac{D_n(\ell,0)}{n}w_\ell - \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell\right|. \quad (6.8.7)$$

To show that the second term goes to zero in probability, consider the characteristic $\chi(t) = \sum_{\ell=0}^{\infty} w_\ell \mathbb{1}\{\xi_{f_1}(t)=\ell\}$. By Lemma 6.8.2, $w_\ell \leq C(\ell+1)$ and hence, $\chi \in \mathscr{C}$. Thus, by Lemma 6.5.8 (i),

$$\left|\sum_{\ell=0}^{\infty}\frac{D_n(\ell,0)}{n}w_\ell - \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell\right| \xrightarrow{P} 0 \quad \text{as } n\to\infty. \quad (6.8.8)$$

To show that the first term in the bound (6.8.7) goes to zero in probability, take any $\epsilon > 0$. Recalling $\sum_{\ell=0}^{\infty} D_n(\ell,0) = n$ and $\sum_{\ell=0}^{\infty}(\ell+1)D_n(\ell,0) = 2n-1$, and taking $t = \eta\log n + a$ for any $n \geq e^{(t(\epsilon)-a)/\eta}$ in Lemma 6.8.3, we obtain

$$\sum_{\ell=0}^{\infty}\frac{D_n(\ell,0)}{n}\left|\frac{\lambda_\ell^\phi(\eta\log n+a)}{n^{\eta\lambda_1^*}} - w_\ell c_\phi e^{\lambda_1^* a}\right| \leq \frac{C'' e^{\lambda_1^* a}\epsilon}{n}\sum_{\ell=0}^{\infty}(\ell+1)D_n(\ell,0) \leq 2C'' e^{\lambda_1^* a}\epsilon.$$

As $\epsilon > 0$ is arbitrary, this shows that the first term in (6.8.7) converges to zero as $n\to\infty$ and completes the proof of the lemma.

■

Define $m^\star := \int_0^\infty u e^{-\lambda_1^* u}\mu_{f_1}(du)$.

**Corollary 6.8.5.** *Fix any $\eta > 0$. Then*

$$\frac{1}{n^{1+\eta\lambda_1^*}}\sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell(\eta\log n) \xrightarrow{P} \frac{1}{\lambda_1^* m^\star}\sum_{\ell=0}^{\infty} p_\ell^0 w_\ell$$

*and for each $k \geq 0$,*

$$\frac{1}{n^{1+\eta\lambda_1^*}}\sum_{\ell=0}^{\infty} D_n(\ell,0)\lambda_\ell^{(k)}(\eta\log n) \xrightarrow{P} \frac{1}{\lambda_1^* m^\star}p_k^1\sum_{\ell=0}^{\infty} p_\ell^0 w_\ell$$

*as $n\to\infty$.*

*Proof.* Follows from Lemma 6.8.4 upon noting that

$$\lambda_\ell(t) = 1 + \int_0^t m_{f_1}(t-s)\mu_{f_1}^{(\ell)}(ds), \quad \lambda_\ell^{(k)}(t) = \mathbb{P}\left(\xi_{f_1}^{(l)}(t) = k-\ell\right) + \int_0^t m_{f_1}^{(k)}(t-s)\mu_{f_1}^{(\ell)}(ds)$$

and observing by Lemma 6.5.8 (ii)

$$\lim_{t\to\infty} e^{-\lambda_1^* t} m_{f_1}(t) = \frac{1}{\lambda_1^* m^\star}, \quad \lim_{t\to\infty} e^{-\lambda_1^* t} m_{f_1}^{(k)}(t) = \frac{p_k^1}{\lambda_1^* m^\star}. \tag{6.8.9}$$

■

**Lemma 6.8.6.** *There exists $\eta_0 > 0$ such that for any $\eta \le \eta_0$, the following limits hold as $n \to \infty$:*

*(i)* $\frac{1}{n^{1+\eta\lambda_1^*}} Z_n(\eta \log n) \xrightarrow{P} \frac{1}{\lambda_1^* m^\star} \sum_{\ell=0}^\infty p_\ell^0 w_\ell$,

*(ii) For any $k \ge 0$,* $\frac{1}{n^{1+\eta\lambda_1^*}} D_n(k, \eta \log n) \xrightarrow{P} \frac{p_k^1}{\lambda_1^* m^\star} \sum_{\ell=0}^\infty p_\ell^0 w_\ell$.

*Proof.* (i) and (ii) follow from (6.8.2) and (6.8.1) respectively along with Corollary 6.8.5. ■

**Corollary 6.8.7.** $\sum_{\ell=0}^\infty p_\ell^1 w_\ell = \lambda_1^* m^\star$.

*Proof.* Note that Lemma 6.8.6 (i) holds in the special case where $f_0 = f_1$ (the model without change point). In this case, $p_\ell^0 = p_\ell^1$ for all $\ell \ge 0$. By Lemma 6.5.8 (ii),

$$\frac{Z_n(\eta_0 \log n)}{e^{\lambda_1^*(T_n + \eta_0 \log n)}} \xrightarrow{a.s.} \frac{W_\infty}{\lambda_1^* m^\star}.$$

Moreover, as $Z(T_n) = n$, therefore, applying Lemma 6.5.8 (ii) again,

$$\frac{e^{\lambda_1^* T_n}}{n} = \frac{1}{e^{-\lambda_1^* T_n} Z(T_n)} \xrightarrow{a.s.} \frac{\lambda_1^* m^\star}{W_\infty}.$$

Using these observations, we obtain

$$\frac{1}{n^{1+\eta_0\lambda_1^*}} Z_n(\eta_0 \log n) = \frac{e^{\lambda_1^* T_n}}{n} \frac{Z_n(\eta_0 \log n)}{e^{\lambda_1^*(T_n + \eta_0 \log n)}} \xrightarrow{a.s.} 1.$$

Comparing this with Lemma 6.8.6 (i) with $f_0 = f_1$ gives the result.

■

Recall that for any $k \geq 0$, $\xi_{f_1}^{(k)}(\cdot)$ is the point process denoting the distribution of birth times of children of a vertex which is of degree $k$ at time zero. The following lemma gives an estimate on the second moment of $\xi_{f_1}^{(k)}(t)$ under Assumption 6.3.1.

**Lemma 6.8.8.** *There exists $C > 0$ and $\beta' < \lambda_1^*$ such that for any $k \geq 0$, $t \geq 0$,*

$$\mathbb{E}\left(\xi_{f_1}^{(k)}(t)\right)^2 \leq C(k+1)^2 e^{2\beta' t}.$$

*Proof.* By Assumption 6.3.1 and Lemma 6.8.1, for any $\beta' \in (\beta_1, \lambda_1^*)$, there exists $\ell_0 \geq 0$ such that for all $\ell \geq \ell_0$, $f_1(\ell) \leq \beta'\ell$. Let $m = \max_{\ell \leq \ell_0} f_1(\ell)$. It is clear that $\xi_{f_1}^{(k)}(\cdot)$ is stochastically dominated by the offspring distribution of a continuous time branching process with attachment function $f^*(\ell) = \beta'\ell + 1 + (m + \beta' k), \ell \geq 0$, which we denote by $\xi_{f^*}^{(k)}(\cdot)$. Applying the second moment obtained in Lemma 6.5.5 (with $\nu = \beta'$ and $\kappa = 1 + m + \beta'(k-1)$) the lemma follows.

■

For $j \geq 0$, $\eta > 0$, let $D_n(k, j, \eta)$ denote the number of vertices of degree $k$ at time $(j+1)\eta \log n$ that were born before time $j\eta \log n$.

**Lemma 6.8.9.** *For any $\eta > 0$, $j \geq 0$, as $n \to \infty$,*

$$\frac{\sum_{k=0}^{\infty}(k+1)D_n(k, j, \eta)}{Z_n(j\eta \log n) n^{\lambda_1^* \eta}} \xrightarrow{P} 0.$$

*Proof.* We will condition on $\mathscr{F}_n(j\eta \log n)$ throughout the proof. Denoting by $\{\xi_{f_1,m}^{(\ell)}(t)\}_{1 \leq m \leq D_n(\ell, j\eta \log n)}$ the degree at time $t + j\eta \log n$ of the m-th vertex of degree $\ell$ at time $j\eta \log n$, observe that

$$\begin{aligned}
\sum_{k=0}^{\infty}(k+1)D_n(k, j, \eta) &= \sum_{k=0}^{\infty}(k+1)\sum_{\ell=0}^{k}\sum_{m=1}^{D_n(\ell, j\eta \log n)} \mathbb{1}\left\{\xi_{f_1,m}^{(\ell)}(\eta \log n) = k - \ell\right\} \\
&= \sum_{\ell=0}^{\infty}\sum_{m=1}^{D_n(\ell, j\eta \log n)}\sum_{k=\ell}^{\infty}(k+1)\mathbb{1}\left\{\xi_{f_1,m}^{(\ell)}(\eta \log n) = k - \ell\right\} = \sum_{\ell=0}^{\infty}\sum_{m=1}^{D_n(\ell, j\eta \log n)}\left(\ell + 1 + \xi_{f_1,m}^{(\ell)}(\eta \log n)\right) \\
&= \sum_{\ell=0}^{\infty}(\ell+1)D_n(\ell, j\eta \log n) + \sum_{\ell=0}^{\infty}\sum_{m=1}^{D_n(\ell, j\eta \log n)}\xi_{f_1,m}^{(\ell)}(\eta \log n) \\
&= 2Z_n(j\eta \log n) - 1 + \sum_{\ell=0}^{\infty}\sum_{m=1}^{D_n(\ell, j\eta \log n)}\xi_{f_1,m}^{(\ell)}(\eta \log n).
\end{aligned}$$

Thus, it suffices to show that as $n \to \infty$,

$$\frac{1}{Z_n(j\eta\log n)} \sum_{\ell=0}^{\infty} \sum_{m=1}^{D_n(\ell, j\eta\log n)} \frac{1}{n^{\lambda_1^*\eta}} \xi_{f_1,m}^{(\ell)}(\eta\log n) \xrightarrow{P} 0. \tag{6.8.10}$$

Note that using Lemma 6.8.8,

$$\begin{aligned}
&\mathrm{Var}\left( \frac{1}{Z_n(j\eta\log n)} \sum_{\ell=0}^{\infty} \sum_{m=1}^{D_n(\ell, j\eta\log n)} \frac{1}{n^{\lambda_1^*\eta}} \xi_{f_1,m}^{(\ell)}(\eta\log n) \right) \\
&\le \frac{1}{Z^2(j\eta\log n) n^{2\lambda_1^*\eta}} \sum_{\ell=0}^{\infty} \sum_{m=1}^{D_n(\ell, j\eta\log n)} \mathbb{E}\left( \xi_{f_1,m}^{(\ell)}(\eta\log n) \right)^2 \le \frac{Cn^{2\beta'\eta}}{Z^2(j\eta\log n) n^{2\lambda_1^*\eta}} \sum_{\ell=0}^{\infty} (\ell+1)^2 D_n(\ell, j\eta\log n).
\end{aligned}$$

Denoting the maximum out-degree at time $j\eta\log n$ of the branching process by $D^{\max}$, note that $D^{\max} + 1 \le Z_n(j\eta\log n)$ and hence,

$$\sum_{\ell=0}^{\infty} (\ell+1)^2 D_n(\ell, j\eta\log n) \le (D^{\max}+1) \sum_{\ell=0}^{\infty} (\ell+1) D_n(\ell, j\eta\log n) \le Z_n(j\eta\log n)(2Z_n(j\eta\log n) - 1).$$

Using this in the above variance bound, we get

$$\mathrm{Var}\left( \frac{1}{Z_n(j\eta\log n)} \sum_{\ell=0}^{\infty} \sum_{m=1}^{D_n(\ell, j\eta\log n)} \frac{1}{n^{\lambda_1^*\eta}} \xi_{f_1,m}^{(\ell)}(\eta\log n) \right) \le \frac{2Cn^{2\beta'\eta} Z^2(j\eta\log n)}{Z^2(j\eta\log n) n^{2\lambda_1^*\eta}} = \frac{2C}{n^{2(\lambda_1^*-\beta')\eta}} \to 0$$

as $n \to \infty$ and hence,

$$\frac{1}{Z_n(j\eta\log n)} \sum_{\ell=0}^{\infty} \sum_{m=1}^{D_n(\ell, j\eta\log n)} \frac{1}{n^{\lambda_1^*\eta}} \xi_{f_1,m}^{(\ell)}(\eta\log n) - \frac{1}{Z_n(j\eta\log n)} \sum_{\ell=0}^{\infty} \sum_{m=1}^{D_n(\ell, j\eta\log n)} \frac{1}{n^{\lambda_1^*\eta}} \mathbb{E}\left( \xi_{f_1,m}^{(\ell)}(\eta\log n) \right) \xrightarrow{P} 0. \tag{6.8.11}$$

By Lemma 6.8.2, we obtain $\beta \in (\lambda_1^* - 1, \lambda_1^*)$ such that $w_\ell(\beta) = \int_0^\infty e^{-\beta s} \mu_{f_1}^{(\ell)}(ds) \le C(\beta)(\ell+1)$. This implies for any $m, \ell$,

$$\mathbb{E}\left( \xi_{f_1,m}^{(\ell)}(\eta\log n) \right) \le C(\beta) n^{\beta\eta} (\ell+1)$$

and consequently,

$$\begin{aligned}
&\frac{1}{Z_n(j\eta\log n)} \sum_{\ell=0}^{\infty} \sum_{m=1}^{D_n(\ell, j\eta\log n)} \frac{1}{n^{\lambda_1^*\eta}} \mathbb{E}\left( \xi_{f_1,m}^{(\ell)}(\eta\log n) \right) \\
&\qquad \le \frac{1}{n^{(\lambda_1^*-\beta)\eta}} \frac{C(\beta)}{Z_n(j\eta\log n)} \sum_{\ell=0}^{\infty} (\ell+1) D_n(\ell, j\eta\log n) \le \frac{2C(\beta)}{n^{(\lambda_1^*-\beta)\eta}} \to 0
\end{aligned} \tag{6.8.12}$$

as $n \to \infty$. From (6.8.11) and (6.8.12), the proof of (6.8.10), and hence the lemma, is complete.

■

**Lemma 6.8.10.** *Let $\phi \in \mathscr{C}$ such that $\lim_{t\to\infty} e^{-\lambda_1^* t} m^\phi(t) = c_\phi$. For $\ell \geq 0$, let $\lambda_\ell^\phi(\cdot)$ be defined as in* (6.8.5). *Fix any $j \geq 0$. There exists $\eta_0 > 0$ such that for any $\eta \leq \eta_0$ and any $a \in \mathbb{R}$, the following limit holds as $n \to \infty$:*

$$\frac{1}{n^{1+(j\eta_0+\eta)\lambda_1^*}} \sum_{\ell=0}^{\infty} D_n(\ell, j\eta_0 \log n) \lambda_\ell^\phi(\eta \log n + a) \xrightarrow{P} c_\phi e^{\lambda_1^* a} \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell.$$

*Proof.* We will proceed by induction. Suppose we can show that for some $j \geq 0$, the assertion of the lemma holds for all $j' \leq j$. Taking $\phi(t) = \mathbb{1}\{t \geq 0\}$ and $\eta = \eta_0$ and recalling $\lim_{t\to\infty} e^{-\lambda_1^* t} m_{f_1}(t) = \frac{1}{\lambda_1^* m^\star}$, we obtain for any $j' \leq j$ and $a \in \mathbb{R}$,

$$\frac{1}{n^{1+(j'+1)\eta_0\lambda_1^*}} Z_n((j'+1)\eta_0 \log n + a) \xrightarrow{P} \frac{1}{\lambda_1^* m^\star} e^{\lambda_1^* a} \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell. \tag{6.8.13}$$

Fix any $\phi \in \mathscr{C}$. Note that for any $\eta \leq \eta_0$,

$$\begin{aligned}
&\left| \frac{1}{n^{1+((j+1)\eta_0+\eta)\lambda_1^*}} \sum_{\ell=0}^{\infty} D_n(\ell, (j+1)\eta_0 \log n) \lambda_\ell^\phi(\eta \log n + a) - c_\phi e^{\lambda_1^* a} \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell \right| \\
&\qquad \leq \sum_{\ell=0}^{\infty} \frac{D_n(\ell, (j+1)\eta_0 \log n)}{n^{1+(j+1)\eta_0\lambda_1^*}} \left| \frac{\lambda_\ell^\phi(\eta \log n + a)}{n^{\eta\lambda_1^*}} - c_\phi e^{\lambda_1^* a} w_\ell \right| \\
&\qquad\qquad + c_\phi e^{\lambda_1^* a} \left| \sum_{\ell=0}^{\infty} \frac{D_n(\ell, (j+1)\eta_0 \log n)}{n^{1+(j+1)\eta_0\lambda_1^*}} w_\ell - \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell \right|. \qquad (6.8.14)
\end{aligned}$$

For any $\epsilon > 0$, by Lemma 6.8.3, there exists $n_0$ such that for all $n \geq n_0$, $\left| \frac{\lambda_\ell^\phi(\eta \log n + a)}{n^{\eta\lambda_1^*}} - c_\phi e^{\lambda_1^* a} w_\ell \right| \leq C'' e^{\lambda_1^* a} \epsilon(\ell+1)$ and hence,

$$\begin{aligned}
&\sum_{\ell=0}^{\infty} \frac{D_n(\ell, (j+1)\eta_0 \log n)}{n^{1+(j+1)\eta_0\lambda_1^*}} \left| \frac{\lambda_\ell^\phi(\eta \log n + a)}{n^{\eta\lambda_1^*}} - c_\phi e^{\lambda_1^* a} w_\ell \right| \\
&\qquad \leq C'' e^{\lambda_1^* a} \epsilon \sum_{\ell=0}^{\infty} \frac{(\ell+1) D_n(\ell, (j+1)\eta_0 \log n)}{n^{1+(j+1)\eta_0\lambda_1^*}} \leq 2C'' e^{\lambda_1^* a} \epsilon \frac{Z_n((j+1)\eta_0 \log n)}{n^{1+(j+1)\eta_0\lambda_1^*}}.
\end{aligned}$$

Therefore, using (6.8.13), the first term in the bound (6.8.14) converges to zero in probability. To estimate the second term in (6.8.14), consider the characteristic $\chi(t) = \sum_{\ell=0}^{\infty} w_\ell \mathbb{1}\{\xi_{f_1}(t) = \ell\}$ and note that by Lemma 6.8.2, $\chi \in \mathscr{C}$. Recall $Z_n^\chi$ from Section 6.6 with $\mathscr{F}_n(0)$ replaced by $\mathscr{F}_n(j\eta_0 \log n)$ and time starting at $T_n + j\eta_0 \log n$. As $Z_n^\chi$ denotes the aggregate $\chi$-score of the children of all vertices born in the interval

$[j\eta_0 \log n, (j+1)\eta_0 \log n]$,

$$\frac{1}{n^{1+(j+1)\eta_0\lambda_1^*}}\left|\sum_{\ell=0}^{\infty} D_n(\ell,(j+1)\eta_0\log n)\, w_\ell - Z_n^\chi\right| \le \frac{C(\lambda_1^*)}{n^{1+(j+1)\eta_0\lambda_1^*}} \sum_{\ell=0}^{\infty}(\ell+1)D_n(\ell,j,\eta_0)$$
$$= \frac{Z_n(j\eta_0\log n)}{n^{1+j\eta_0\lambda_1^*}} \frac{C(\lambda_1^*)}{Z_n(j\eta_0\log n)n^{\eta_0\lambda_1^*}} \sum_{\ell=0}^{\infty}(\ell+1)D_n(\ell,j,\eta_0) \xrightarrow{P} 0 \quad (6.8.15)$$

as $n\to\infty$ by (6.8.13) and Lemma 6.8.9, where $C(\lambda_1^*)$ is the constant appearing in Lemma 6.8.2. By Theorem 6.6.1 (taking $a=\eta_0\log n$) and (6.8.13), if $\eta_0$ is chosen such that $\frac{Ce^{C'\eta_0\log n}}{\sqrt{n}}\to 0$, where $C, C'$ are the constants appearing in Theorem 6.6.1 (note that this condition on $\eta_0$ is independent of $j$),

$$\frac{1}{n^{1+(j+1)\eta_0\lambda_1^*}}\left|Z_n^\chi - \sum_{\ell=0}^{\infty} D_n(\ell, j\eta_0\log n)\lambda_\ell^\chi(\eta_0\log n)\right| \le \frac{Ce^{C'\eta_0\log n}}{n^{1+(j+1)\eta_0\lambda_1^*}}\sqrt{Z_n(j\eta_0\log n)}$$
$$\le \frac{Ce^{C'\eta_0\log n}}{\sqrt{n}}\sqrt{\frac{Z_n(j\eta_0\log n)}{n^{1+(j+1)\eta_0\lambda_1^*}}} \xrightarrow{P} 0 \quad (6.8.16)$$

where we recall $\lambda_\ell^\chi(t) = \int_0^t m_{f_1}^\chi(t-s)\mu_{f_1}^{(\ell)}(ds)$. By (6.8.15) and (6.8.16), we obtain

$$\left|\sum_{\ell=0}^{\infty}\frac{D_n(\ell,(j+1)\eta_0\log n)}{n^{1+(j+1)\eta_0\lambda_1^*}}w_\ell - \sum_{\ell=0}^{\infty}\frac{D_n(\ell,j\eta_0\log n)}{n^{1+(j+1)\eta_0\lambda_1^*}}\lambda_\ell^\chi(\eta_0\log n)\right|$$
$$\le \frac{1}{n^{1+(j+1)\eta_0\lambda_1^*}}\left|\sum_{\ell=0}^{\infty} D_n(\ell,(j+1)\eta_0\log n)\, w_\ell - Z_n^\chi\right|$$
$$+ \frac{1}{n^{1+(j+1)\eta_0\lambda_1^*}}\left|Z_n^\chi - \sum_{\ell=0}^{\infty} D_n(\ell, j\eta_0\log n)\lambda_\ell^\chi(\eta_0\log n)\right| \xrightarrow{P} 0 \quad (6.8.17)$$

Next, we will show that

$$e^{-\lambda_1^* t} m_{f_1}^\chi(t) \to 1 \text{ as } t\to\infty. \quad (6.8.18)$$

To see this, first note that it follows from Assumption 6.2.1 (iii) that there exists $\beta < \lambda_1^*$ such that $\mathbb{E}\left(\xi_{f_1}(t)\right) \le Ce^{\beta t}$. Moreover, $w_\ell \le C(\ell+1)$ for all $\ell \ge 0$. These observations imply

$$\sum_{k=0}^{\infty}\sup_{t\in[k,k+1]}\left[e^{-\lambda_1^* t}E(\chi(t))\right] \le C\sum_{k=0}^{\infty}\sup_{t\in[k,k+1]}\left[e^{-\lambda_1^* t}\sum_{\ell=0}^{\infty}(\ell+1)\,\mathbb{P}\left(\xi_{f_1}(t)=\ell\right)\right]$$
$$= C\sum_{k=0}^{\infty}\sup_{t\in[k,k+1]}\left[e^{-\lambda_1^* t}\mathbb{E}\left(\xi_{f_1}(t)+1\right)\right] \le C'\sum_{k=0}^{\infty}\sup_{t\in[k,k+1]}\left[e^{-\lambda_1^* t}e^{\beta t}\right] \le C'e^{\beta}\sum_{k=0}^{\infty}e^{-(\lambda_1^*-\beta)k} < \infty$$

where $C, C' > 0$ are constants. Thus, by Proposition 2.2 of (Nerman, 1981) and Corollary 6.8.7, it follows that

$$\lim_{t\to\infty} e^{-\lambda_1^* t} m_{f_1}^{\chi}(t) = \frac{1}{\lambda_1^* m^\star} \sum_{\ell=0}^{\infty} w_\ell \lambda_1^* \int_0^\infty e^{-\lambda_1^* s} \mathbb{P}\left(\xi_{f_1}(s) = \ell\right) ds = \frac{1}{\lambda_1^* m^\star} \sum_{\ell=0}^{\infty} w_\ell p_\ell^1 = 1.$$

Using this, the definition of $\lambda_\ell^\chi$, the fact that $\chi \in \mathscr{C}$ and the induction hypothesis, we obtain

$$\frac{1}{n^{1+(j+1)\eta_0\lambda_1^*}} \sum_{\ell=0}^{\infty} D_n(\ell, j\eta_0 \log n) \lambda_\ell^\chi(\eta_0 \log n) \xrightarrow{P} \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell \quad \text{as } n \to \infty. \tag{6.8.19}$$

From (6.8.17) and (6.8.19), we conclude that the second term in the bound (6.8.14) goes to 0 as $n \to \infty$ which proves that

$$\left| \frac{1}{n^{1+((j+1)\eta_0+\eta)\lambda_1^*}} \sum_{\ell=0}^{\infty} D_n(\ell, (j+1)\eta_0 \log n) \lambda_\ell^\phi(\eta \log n + a) - c_\phi e^{\lambda_1^* a} \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell \right| \xrightarrow{P} 0$$

establishing the induction hypothesis for $j+1$. The induction hypothesis for $j = 0$ is true by Lemma 6.8.4. Thus, the lemma is proved.

■

**Lemma 6.8.11.** *For any $k \geq 0, \theta > 0$ and $a \in \mathbb{R}$, as $n \to \infty$:*

$$\frac{1}{n^{1+\theta\lambda_1^*}} Z_n(\theta \log n + a) \xrightarrow{P} \frac{1}{\lambda_1^* m^\star} e^{\lambda_1^* a} \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell, \qquad \frac{D_n(k, \theta \log n + a)}{Z_n(\theta \log n + a)} \xrightarrow{P} p_k^1.$$

*Proof.* The first assertion follows by the exact argument used to derive (6.8.13).

To prove the second assertion, fix any $k \geq 0$. Obtain $\eta_0 > 0$ as in Lemma 6.8.10. Moreover, without loss of generality, assume $\eta_0$ is small enough so that $\frac{Ce^{C'\eta_0 \log n}}{\epsilon^2} \frac{1}{n^{\omega - \widetilde{\theta} - \frac{1}{2}}} \to 0$, where $C, C', \omega, \widetilde{\theta}$ are as in Lemma 6.7.11. Let $j \geq 0$, $\eta \in [0, \eta_0)$ such that $\theta = j\eta_0 + \eta$. Recall that the probability bound obtained in Lemma 6.7.11 conditionally on $\mathscr{F}_n(0)$ was in terms of deterministic constants and $n$, the total number of vertices at time 0. Thus, replacing $\mathscr{F}_n(0)$ by $\mathscr{F}_n(j\eta_0 \log n)$ and time starting from $T_n + j\eta_0 \log n$, Lemma 6.7.11 implies

$$\frac{1}{Z_n(j\eta_0 \log n)} D_n(k, \theta \log n + a) - \frac{1}{Z_n(j\eta_0 \log n)} \sum_{\ell=0}^{\infty} D_n(\ell, j\eta_0 \log n) \lambda_\ell^{(k)}(\eta \log n + a) \xrightarrow{P} 0, \quad \text{as } n \to \infty.$$

From Lemma 6.8.10 (taking $\phi(t) = \mathbb{1}\{t \geq 0\}$), $\frac{Z_n(j\eta_0 \log n)}{Z_n(\theta \log n + a)} \xrightarrow{P} 0$ if $\eta > 0$ and $\frac{Z_n(j\eta_0 \log n)}{Z_n(\theta \log n + a)} \xrightarrow{P} e^{-\lambda_1^* a}$ if $\eta = 0$ and thus, multiplying both sides of the above by $\frac{Z_n(j\eta_0 \log n)}{Z_n(\theta \log n + a)}$, we obtain

$$\frac{D_n(k, \theta \log n + a)}{Z_n(\theta \log n + a)} - \frac{1}{Z_n(\theta \log n + a)} \sum_{\ell=0}^{\infty} D_n(\ell, j\eta_0 \log n) \lambda_\ell^{(k)}(\eta \log n + a) \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{6.8.20}$$

Taking $\phi(t) = \mathbb{1}\{\xi_{f_1}(t) = k\}$, we see that $\lambda_\ell^{\phi} = \lambda_\ell^{(k)}$ for each $\ell \geq 0$. Moreover, recall from (6.8.9)

$$\lim_{t \to \infty} e^{-\lambda_1^* t} m_{f_1}^{(k)}(t) = \frac{p_k^1}{\lambda_1^* m^\star}.$$

Thus, from Lemma 6.8.10,

$$\frac{1}{n^{1+\theta\lambda_1^*}} \sum_{\ell=0}^{\infty} D_n(\ell, j\eta_0 \log n) \lambda_\ell^{(k)}(\eta \log n + a) \xrightarrow{P} \frac{p_k^1}{\lambda_1^* m^\star} e^{\lambda_1^* a} \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell. \tag{6.8.21}$$

Using (6.8.21) and the first assertion of the lemma in (6.8.20), the second assertion follows.

■

Define $a_0 := \frac{1}{\lambda_1^*} \log\left(\frac{\lambda_1^* m^\star}{\sum_{\ell=0}^{\infty} p_\ell^0 w_\ell}\right)$. Also, let $T_n^\theta := T_{n^{1+\lambda_1^*\theta}}$ denote the first time the branching process has $n^{1+\lambda_1^*\theta}$ vertices.

**Lemma 6.8.12.** $T_n^\theta - \theta \log n \xrightarrow{P} a_0$.

*Proof.* Follows immediately from the first assertion of Lemma 6.8.11. ■

**Theorem 6.8.13.** *For any $k \geq 0$, $\theta > 0$, as $n \to \infty$,*

$$\frac{D_n(k, T_n^\theta)}{n^{1+\lambda_1^*\theta}} \xrightarrow{P} p_k^1.$$

*Proof.* In the proof, we will abbreviate $z^* = \frac{1}{\lambda_1^* m^\star} \sum_{\ell=0}^{\infty} p_\ell^0 w_\ell$. Fix any $k \geq 0$, $\theta > 0$. Take any $\epsilon \in (0,1)$. By the same argument as in the proof of Lemma 6.7.8,

$$\sup_{t \leq 2\epsilon} |D_n(k, \theta \log n + a_0 - \epsilon + t) - D_n(k, \theta \log n + a_0 - \epsilon)| \leq \left(Z_n(\theta \log n + a_0 + \epsilon) - Z_n(\theta \log n + a_0 - \epsilon)\right) + Y_n. \tag{6.8.22}$$

where, conditionally on $\mathscr{F}_n(\theta\log n + a_0 - \epsilon)$, $Y_n$ is distributed as $\sum_{\ell=0}^{k} \mathrm{Bin}\left(D_n\left(\ell, \theta\log n + a_0 - \epsilon\right), q_\ell(2\epsilon)\right)$. Observe that by the first assertion in Lemma 6.8.11, for small enough $\epsilon$,

$$\frac{Z_n(\theta\log n + a_0 + \epsilon) - Z_n(\theta\log n + a_0 - \epsilon)}{n^{1+\lambda_1^*\theta}} \xrightarrow{P} e^{\lambda_1^*\epsilon} - e^{-\lambda_1^*\epsilon} \le 4\lambda_1^*\epsilon. \tag{6.8.23}$$

Note that for any $C > 0$,

$$\begin{aligned}\mathbb{P}\left(Y_n > C\sqrt{\epsilon}n^{1+\lambda_1^*\theta}\right) \le \mathbb{P}\left(Y_n > C\sqrt{\epsilon}n^{1+\lambda_1^*\theta}, Z_n(\theta\log n + a_0 - \epsilon) \le \epsilon^{-1/2}n^{1+\lambda_1^*\theta}\right) \\ + \mathbb{P}\left(Z_n(\theta\log n + a_0 - \epsilon) > \epsilon^{-1/2}n^{1+\lambda_1^*\theta}\right).\end{aligned} \tag{6.8.24}$$

For $\epsilon$ sufficiently small, by the first assertion of Lemma 6.8.11, as $n \to \infty$,

$$\mathbb{P}\left(Z_n(\theta\log n + a_0 - \epsilon) > \epsilon^{-1/2}n^{1+\lambda_1^*\theta}\right) \to 0. \tag{6.8.25}$$

Let $\mathscr{H}_n := \mathscr{F}_n(\theta\log n + a_0 - \epsilon)$. Using Lemma 6.7.3,

$$\begin{aligned}\mathbb{E}(Y_n \mid \mathscr{H}_n) = \sum_{\ell=0}^{k} D_n\left(\ell, \theta\log n + a_0 - \epsilon\right) q_\ell(2\epsilon) \le C'\epsilon \sum_{\ell=0}^{k} (\ell+1) D_n\left(\ell, \theta\log n + a_0 - \epsilon\right) \\ \le 2C'\epsilon Z_n(\theta\log n + a_0 - \epsilon).\end{aligned}$$

Thus, choosing $C > 4C'$, using Chebychev's inequality, conditionally on $\mathscr{H}_n$ on the event $\{Z_n(\theta\log n + a_0 - \epsilon) \le \epsilon^{-1/2}n^{1+\lambda_1^*\theta}\}$,

$$\begin{aligned}\mathbb{P}\left(Y_n > C\sqrt{\epsilon}n^{1+\lambda_1^*\theta} \mid \mathscr{H}_n\right) &\le \mathbb{P}\left(Y_n - \mathbb{E}(Y_n \mid \mathscr{H}_n) > \frac{C}{2}\sqrt{\epsilon}n^{1+\lambda_1^*\theta} \mid \mathscr{H}_n\right) \\ &\le \frac{4\,\mathrm{Var}(Y_n \mid \mathscr{H}_n)}{C^2\epsilon n^{2(1+\lambda_1^*\theta)}} = \frac{4\sum_{\ell=0}^{k} D_n\left(\ell, \theta\log n + a_0 - \epsilon\right) q_\ell(2\epsilon)\left(1 - q_\ell(2\epsilon)\right)}{C^2\epsilon n^{2(1+\lambda_1^*\theta)}} \\ &\le \frac{4C'\epsilon\sum_{\ell=0}^{k}(\ell+1) D_n\left(\ell, \theta\log n + a_0 - \epsilon\right)}{C^2\epsilon n^{2(1+\lambda_1^*\theta)}} \le \frac{8C' Z_n(\theta\log n + a_0 - \epsilon)}{C^2 n^{2(1+\lambda_1^*\theta)}} \\ &\le \frac{8C'}{C^2\sqrt{\epsilon}n^{1+\lambda_1^*\theta}} \to 0 \quad \text{as } n \to \infty.\end{aligned} \tag{6.8.26}$$

Using (6.8.25) and (6.8.26) in (6.8.24), we conclude

$$\mathbb{P}\left(Y_n > C\sqrt{\epsilon}n^{1+\lambda_1^*\theta}\right) \to 0 \quad \text{as } n \to \infty. \tag{6.8.27}$$

Using (6.8.23), (6.8.27) and (6.8.22), we conclude that there exist $C_0 > 0, \epsilon_0 > 0$ such that for all $\epsilon \in (0, \epsilon_0)$,

$$\mathbb{P}\left(\sup_{t\le 2\epsilon} |D_n(k, \theta\log n + a_0 - \epsilon + t) - D_n(k, \theta\log n + a_0 - \epsilon)| > C_0\sqrt{\epsilon}n^{1+\lambda_1^*\theta}\right) \to 0 \quad \text{as } n \to \infty. \tag{6.8.28}$$

From (6.8.28) and Lemma 6.8.12, as $n \to \infty$,

$$\begin{aligned}\mathbb{P}\left(|D_n(k, T_n^\theta) - D_n(k, \theta\log n + a_0 - \epsilon)| > C_0\sqrt{\epsilon}n^{1+\lambda_1^*\theta}\right) &\le \mathbb{P}\left(\left|T_n^\theta - \theta\log n - a_0\right| > 2\epsilon\right) \\ &+ \mathbb{P}\left(\sup_{t\le 2\epsilon} |D_n(k, \theta\log n + a_0 - \epsilon + t) - D_n(k, \theta\log n + a_0 - \epsilon)| > C_0\sqrt{\epsilon}n^{1+\lambda_1^*\theta}\right) \to 0. \end{aligned} \tag{6.8.29}$$

For any $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{D_n(k, T_n^\theta)}{n^{1+\lambda_1^*\theta}} - p_k^1\right| > 2C_0\sqrt{\epsilon}\right) &\le \mathbb{P}\left(\left|\frac{D_n(k, T_n^\theta)}{n^{1+\lambda_1^*\theta}} - \frac{D_n(k, \theta\log n + a_0 - \epsilon)}{n^{1+\lambda_1^*\theta}}\right| > C_0\sqrt{\epsilon}\right) \\ &+ \mathbb{P}\left(\left|\frac{D_n(k, \theta\log n + a_0 - \epsilon)}{n^{1+\lambda_1^*\theta}} - p_k^1\right| > C_0\sqrt{\epsilon}\right). \end{aligned} \tag{6.8.30}$$

By Lemma 6.8.11,

$$\frac{D_n(k, \theta\log n + a_0 - \epsilon)}{n^{1+\lambda_1^*\theta}} = \frac{D_n(k, \theta\log n + a_0 - \epsilon)}{Z_n(\theta\log n + a_0 - \epsilon)} \frac{Z_n(\theta\log n + a_0 - \epsilon)}{n^{1+\lambda_1^*\theta}} \xrightarrow{P} p_k^1 e^{-\lambda_1^*\epsilon},$$

and therefore, there is $\epsilon_1 \le \epsilon_0$ such that for all $\epsilon \in (0, \epsilon_1)$,

$$\left|\frac{D_n(k, \theta\log n + a_0 - \epsilon)}{n^{1+\lambda_1^*\theta}} - p_k^1\right| \xrightarrow{P} p_k^1(1 - e^{-\lambda_1^*\epsilon}) \le p_k^1\lambda_1^*\epsilon < C_0\sqrt{\epsilon}. \tag{6.8.31}$$

For $\epsilon \in (0, \epsilon_1)$, using (6.8.29) and (6.8.31) in (6.8.30), we conclude

$$\mathbb{P}\left(\left|\frac{D_n(k, T_n^\theta)}{n^{1+\lambda_1^*\theta}} - p_k^1\right| > 2C_0\sqrt{\epsilon}\right) \to 0 \quad \text{as } n \to \infty$$

proving the theorem.

■

### 6.8.2 Proof of Theorem 6.3.16:

We will prove (a) of the Theorem. The remaining results follow via straightforward modifications of the arguments for (a). For (a) recall that we first grow the tree using the uniform attachment scheme with $f_0 \equiv 1$ till it is of size $n^\gamma$ and then use the preferential attachment scheme. We will assume that $\mathscr{T}_n^{\boldsymbol{\theta}}$ has been constructed as follows:

(a) Generate the genealogical tree according to a rate one Yule process $\{\mathscr{T}^{\text{Yule}}(t) : t \geq 0\}$ as in Definition 6.5.3 run for ever.

(b) To obtain $\mathscr{T}_n^{\boldsymbol{\theta}}$, let $\mathscr{T}_{n^\gamma} = \mathscr{T}^{\text{Yule}}(T_{n^\gamma})$. Now every vertex in $\mathscr{T}_{n^\gamma}$ switches to offspring dynamics giving birth to children at rate corresponding to the number of children $+1+\alpha$ (thus modulated by the function $f_1$). Write $\text{BP}_n(\cdot)$ for the combined process and stop this process at time $T_n$ and let $\mathscr{T}_n^{\boldsymbol{\theta}} = \text{BP}_n(T_n)$.

The following describes asymptotics for the above continuous time construction.

**Proposition 6.8.14.** *For the process* $\text{BP}_n(\cdot)$ *as constructed above:*

(a) *The stopping time* $T_{n^\gamma}$ *satisfies,*

$$T_{n^\gamma} - \gamma \log n \xrightarrow{\text{a.e.}} \tilde{W},$$

*where* $\tilde{W} = -\log W$ *and* $W = \exp(1)$.

(b) *Let* $\omega_n \to \infty$ *arbitrarily slowly. Then there exists a constant* $C > 0$ *independent of* $\omega_n$ *such that*

$$\mathbb{P}\left(\sup_{t\geq 0}\left|\frac{e^{-(2+\alpha)t}|\text{BP}_n(t+T_{n^\gamma})|}{n^\gamma} - 1\right| > \frac{\omega_n}{n^{\gamma/2}}\right) \leq \frac{C}{\omega_n^2}.$$

*In particular whp as* $n \to \infty$,

$$\left|T_n - \frac{1-\gamma}{2+\alpha}\log n\right| \leq \frac{\omega_n}{n^{\gamma/2}}.$$

*Proof.* Part(a) follows from Lemma 6.5.4. To prove (b), recall that for $t > T_{n^\gamma}$, all individuals switch to offspring dynamics modulated by $f_1$. For the rest of the proof, we proceed conditional on the history of

the process till time $T_{n^\gamma}$. Using Proposition 6.5.7,

$$M_1(t) := \left(e^{-(2+\alpha)t}|BP_n(t+T_{n^\gamma})| - n^\gamma\right) + \frac{1-e^{-(2+\alpha)t}}{(2+\alpha)}, \qquad t \geq 0,$$

and

$$M_2(t) := e^{-2(2+\alpha)t}|BP_n(t+T_{n^\gamma})|^2 - \int_0^t \alpha e^{-2(2+\alpha)s}|BP_n(s+T_{n^\gamma})|ds - \frac{e^{-2(2+\alpha)t}}{2(2+\alpha)}, \qquad t \geq 0,$$

are martingales. Using these expressions, it can be deduced that

$$\sup_{t\geq 0}\mathbb{E}\left(M_1^2(t)\right) \leq Cn^\gamma$$

for some constant $C > 0$. An appeal to Doob's $\mathbb{L}^2$-maximal inequality then proves the first assertion of Proposition 6.8.14(b) which then results in the second assertion.

■

Fix constant $B$ and a sequence $\omega_n = o(n^{\gamma/2}) \uparrow \infty$ and consider the following construction $\tilde{\mathcal{T}}_n^+(B,\omega_n)$ related to the above continuous time construction of $\mathcal{T}_n^{\boldsymbol{\theta}}$:

(a) Run a rate one Yule process for time $\gamma \log n + B$.

(b) Now every vertex in the Yule process switches dynamics so that it reproduces at rate equal to the number of children $+1+\alpha$. Grow this process for **an additional** time $t_n^+ := \frac{1-\gamma}{2+\alpha}\log n + \frac{\omega_n}{n^{\gamma/2}}$.

Analogously define $\tilde{\mathcal{T}}_n^-(B,\omega_n)$ where in the above construction we wait till time $\log n - B$ before switching dynamics and run the new dynamics for time $t_n^- := \frac{1-\gamma}{2+\alpha}\log n - \frac{\omega_n}{n^{\gamma/2}}$. By Proposition 6.8.14 given any $\varepsilon > 0$ we can choose a constant $B = B(\varepsilon)$ such that for any $\omega_n \uparrow \infty$, we can produce a coupling between $\mathcal{T}_n^{\boldsymbol{\theta}}$ and $\tilde{\mathcal{T}}_n^+(B,\omega_n)$ such that for all large $n$, with probability at least $1-\varepsilon$ $\mathcal{T}_n^{\boldsymbol{\theta}} \subseteq \tilde{\mathcal{T}}_n^+(B,\omega_n)$ where we see the object on the left as a subtree of the object on the right with the same root. A similar assertion holds with $\tilde{\mathcal{T}}_n^-(B,\omega_n) \subseteq \mathcal{T}_n^{\boldsymbol{\theta}}$. Using these couplings, the following Proposition completes the proof of Theorem 6.3.16 with part(a) of the Proposition proving the lower bound while part(b) proving the upper bound.

**Proposition 6.8.15.** *Fix $B > 0$ and $\omega_n = o(\log n) \uparrow \infty$.*

(a) *Consider the degree of the root $D_n^-(\rho)$ in $\tilde{\mathcal{T}}_n^-(B,\omega_n)$. Then $D_n^-(\rho) \gg n^{(1-\gamma)/(2+\alpha)}\log n/\omega_n$ whp.*

(b) *Consider the maximal degree* $M_n^+(1)$ *in* $\tilde{\mathscr{T}}_n^+(B,\omega_n)$. *Then* $\exists A > 0$ *such that whp as* $n \to \infty$, $M_n^+(1) \ll An^{(1-\gamma)/(2+\alpha)}(\log n)^2$.

**Proof:** We start with (a). Note that each individual in the original Yule process reproduces according to a rate one Poisson process. In particular standard bounds for a Poisson random variable implies that the degree of the root in $\tilde{\mathscr{T}}_n^-(B,\omega_n)$ by time $\gamma \log n - B$ when the dynamics is switched to preferential attachment dynamics satisfies

$$|\deg_n(\rho, \gamma \log n - B) - \gamma \log n| = O_P(\sqrt{\log n}). \tag{6.8.32}$$

Now let $\{Y_i(\cdot) : i \geq 1\}$ be a collection of independent rate one Yule processes. Comparing rates, the degree of the root after $\gamma \log n - B$ we get that

$$\deg_n(\gamma \log n - B + \cdot) \succeq_{\text{st}} \sum_{i=1}^{\deg_n(\rho, \gamma \log n - B)} Y_i(\cdot), \tag{6.8.33}$$

Using (6.8.32), Lemma 6.5.4 and standard tail bounds for the Geometric distribution now completes the proof.

Let us now prove (b). Recall that after the change point, dynamics are modulated by $f_1(\cdot) := \cdot + 1 + \alpha$. Let $A$ denote the smallest integer $\geq \alpha + 1$. Let $\xi_{f_1}$ be the corresponding continuous time offspring point process. Comparing rates we see that

$$\xi_{f_1}(\cdot) \leq_{st} \sum_{i=1}^{A+2} Y_i(\cdot), \tag{6.8.34}$$

where as before $\{Y_i(\cdot) : i \geq 1\}$ is a collection of independent rate one Yule processes. For every vertex $v$ write $\deg_n(v)$ for the degree of the vertex at time $\log n + B + t_n^+$ when we have finished constructing the process $\tilde{\mathscr{T}}_n^+(B,\omega_n)$. Abusing notation, write $T_v$ for the time of birth of vertex $v$. We will break up the proof of (b) into two cases:

**(b1) Maximal degree for vertices born after** $\log n + B$**:** Define

$$\mathbb{A}_n = \left\{ v \in \tilde{\mathscr{T}}_n^+(B,\omega_n) : T_v \in [\log n + B,\ \log n + B + t_n^+],\ \deg_n(v) > Cn^{\frac{1-\gamma}{2+\alpha}}(\log n)^2. \right\},$$

where $C$ is an appropriate large constant that will be chosen later. The aim is to show that we can choose $C$ such that $\mathbb{E}(|\mathbb{A}_n|) \to 0$, as $n \to \infty$. This would then imply

$$\mathbb{P}(\exists v \in \tilde{\mathcal{T}}_n^+(B,\omega_n), T_v \geq \log n + B \ \deg_n(v) > Cn^{\frac{1-\gamma}{2+\alpha}}(\log n)^2) \to 0. \tag{6.8.35}$$

Let $k_n := Cn^{\frac{1-\gamma}{2+\alpha}}(\log n)^2)$ and let $\tilde{\mathcal{T}}_n^+(t)$ denote the tree at time $t$. Since the offspring distribution of each new vertex born at $t > \log n + B$ is a Yule process then, by Lemma 6.5.4 the probably a new vertex has degree greater than $k_n$ by time $t_n^+$ is given by

$$P(\text{Geom}(e^{t-t_n^+}) \geq k_n) \leq e^{k_n e^{t-t_n^+}}$$

Note that new vertices are produced at rate $(2+\alpha)|\tilde{\mathcal{T}}_n^+(t)|-1$. As in the proof of Proposition 6.8.14 $M(t) := e^{-(2+\alpha)t}|\tilde{\mathcal{T}}_n^+(t)| + \frac{1}{(2+\alpha)}e^{-(2+\alpha)t}$, $t \geq \log n + B$ is a martingale. Noting $\mathbb{E}|\tilde{\mathcal{T}}_n^+(\log n + B)| = e^B n^\gamma$ we get that

$$\mathbb{E}|\tilde{\mathcal{T}}_n^+(t)| = C' n^\gamma e^{(2+\alpha)t} \text{ for } t \geq \log n + B$$

where $C'$ is a constant depending only on $B, \alpha$. Thus

$$\mathbb{E}(|\mathbb{A}_n|) \leq C'' n^\gamma \int_0^{t_n^+} e^{k_n e^{t-t_n^+}} e^{(2+\alpha)t} dt$$

where $C''$ depends only on $B, \alpha$ and it is sufficient to check the following lemma.

**Lemma 6.8.16.** *Let*

$$I_n := n^\gamma \int_0^{t_n^+} e^{-C(\log n)^2 n^{\frac{1-\gamma}{2+\alpha}} e^{t-t_n^+}} e^{(2+\alpha)t} dt \tag{6.8.36}$$

*For sufficiently large $C$, $I_n \to 0$ as $n \to \infty$.*

*Proof.* Writing $a := \frac{1-\gamma}{2+\alpha}$ and $b := 2+\alpha$, algebraic manipulations result in the form:

$$I_n \leq n^\gamma (\log n)^{-2b} e^{b\frac{w_n}{n^{\gamma/2}}} \Gamma\left(b, C(\log n)^2 e^{-\frac{w_n}{n^{\gamma/2}}}\right) := \mathscr{E}_n. \tag{6.8.37}$$

where $\Gamma(b,z) = \int_z^\infty e^{-t}t^{b-1}dt$ is the upper incomplete Gamma function. Known asymptotics for the incomplete Gamma function $\Gamma(b,z) = \Omega(z^{b-1}e^{-z})$ as $z \to \infty$ imply

$$\mathscr{E}_n \sim n^{\gamma - C\log n e^{-\frac{w_n}{n^{\gamma/2}}}} (\log n)^{-2} e^{-\frac{w_n}{n^{\gamma/2}}} \to 0.$$

■

**(b2) Maximal degree for vertices born before** $\log n + B$**:** We prove that vertices born before $\gamma \log n + B$ cannot have too large of a maximal degree in $\tilde{\mathscr{T}}_n^+(B,\omega_n)$. To simplify notation, write the following for the two times:

$$\Delta_n := \gamma \log n + B, \qquad \Upsilon_n := \gamma \log n + B + t_n^+. \tag{6.8.38}$$

Further write $\deg(v,t)$ for the degree of a vertex $v$ at time $t$ with the convention that $\deg(v,t) := 0$ for $t < T_v$. Write $\deg_n(v) := \deg(v,\Upsilon_n)$ for the final degree of $v$ in $\tilde{\mathscr{T}}_n^+(B,\omega_n)$. Finally in the construction of the tree $\tilde{\mathscr{T}}_n^+(B,\omega_n)$, for any $0 \le t \le \Upsilon_n$, write $\tilde{\mathscr{T}}_n^+(t)$ for the tree at time $t$.

Fix $C > 0$ and let $\mathbb{B}_n$ be the set of vertices born before $\log n + B$ whose final degree is too large i.e.

$$\mathbb{B}_n := \{v \in \mathrm{BP}_n : T_v \le \log n + B, \deg_n(v) > C n^{\frac{1-\gamma}{2+\alpha}} (\log n)^2.\}$$

where $\deg_n(v)$ is the degree of vertex $v$ in the final tree $\tilde{\mathscr{T}}_n^+(B,\omega_n)$.

**Proposition 6.8.17.** *We can choose $C < \infty$ such that $\mathbb{P}(\mathbb{B}_n \ge 1) \to 0$ as $n \to \infty$.*

The plan is as follows: we control the maximal degree of vertices born in the early (pre $\Delta_n$) tree then show that none of these early vertices have time to accumulate too many edges in the remaining $\Upsilon_n - \Delta_n$ time period.

*Proof.* Consider the tree $\tilde{\mathscr{T}}_n^+(\Delta_n)$. Let $M_n(\Delta_n) := \max_{v \in \tilde{\mathscr{T}}_n^+(\Delta_n)} deg(v,\Delta_n)$ be the maximal degree of vertices in $\tilde{\mathscr{T}}_n^+(\Delta_n)$ at time $\Delta_n$. Let $\ell_n := 10e\log n$ and fix a sequence $\omega_n \uparrow \infty$. By the union bound,

$$\begin{aligned}\mathbb{P}(\mathbb{B}_n \ge 1) &\le \mathbb{P}(\mathbb{B}_n \ge 1, |\tilde{\mathscr{T}}_n^+(\Delta_n)| < \omega_n n^\gamma, M_n \le \ell_n) \\ &\quad + \mathbb{P}(|\tilde{\mathscr{T}}_n^+(\Delta_n)| \ge \omega_n n^\gamma) + \mathbb{P}(M_n > \ell_n)\end{aligned}$$

Lemmas 6.8.18 and 6.8.19 which bound the three terms on the right complete the proof of the Proposition. ■

**Lemma 6.8.18.** *For $C$ large enough $\mathbb{P}(\mathbb{B}_n \geq 1, |\tilde{\mathcal{T}}_n^+(\Delta_n)| < \omega_n n^\gamma, M_n \leq \ell_n) \to 0$ as $n \to \infty$.*

*Proof.* Let $\mathbb{G}_n = \{|\tilde{\mathcal{T}}_n^+(\Delta_n)| < \omega_n n^\gamma, M_n \leq \ell_n\}$. It is sufficient to show $\mathbb{P}(\mathbb{B}_n \geq 1 | \mathbb{G}_n) \to 0$. Conditional on $\mathbb{G}_n$, we will construct a random variable that stochastically bounds the growth of degrees in the process $\tilde{\mathcal{T}}_n^+(t)$ for $t \geq \Delta_n$. Let $\{X_i(\cdot) : 1 \leq i \leq n^\gamma \omega_n\}$ be a collection of independent rate one Yule processes each starting with $\ell_n + \lceil \alpha \rceil$ individuals at time 0 and run each for time $t_n^+ = \frac{1-\gamma}{2+\alpha} \log n + \frac{\omega_n}{n^{\gamma/2}}$. Consider $\mathscr{M}_n = \max_{1 \leq i \leq \omega_n n^\gamma} X_i(t_n^+)$.

On the event $\mathbb{G}_n$, the degree evolution of $\tilde{\mathcal{T}}_n^+$ after time $\Delta_n$ is as follows: Sample $\tilde{\mathcal{T}}_n^+(\Delta_n)$ conditional on $\mathbb{G}_n$ i.e. the event that there are fewer than $\omega_n n^\gamma$ vertices and the maximal degree is less than $\ell_n$. For each vertex, $v$, in $\tilde{\mathcal{T}}_n^+(\Delta_n)$ we run an independent, rate 1 Yule process starting with $\deg(v, \Delta_n) + \alpha$ individuals for time $t_n^+$. Our new process starts each Yule process as if each individual has maximal degree at time $\gamma \log n + B$. In particular on the event $\mathbb{G}_n$, the maximal degree $M_n(\Upsilon_n)$ at time $\Upsilon_n$ satisfies $M_n(\Upsilon_n) \preceq_{\text{st}} \mathscr{M}_n$. The rest of the proof analyzes $\mathscr{M}_n$. Using the union bound gives,

$$\mathbb{P}\left(\mathbb{B}_n \geq 1 | \mathbb{G}_n\right) \leq \mathbb{P}\left(\mathscr{M}_n \geq C n^{\frac{1-\gamma}{2+\alpha}} (\log n)^2\right) \leq \omega_n n^\gamma \, \mathbb{P}\left(X_i(t_n^+) \geq C n^{\frac{1-\gamma}{2+\alpha}} (\log n)^2\right).$$

Now for a rate one Yule process started with $m$ individuals at time zero say $Y^m(\cdot)$ for fixed $t$, $Y^m(t)$ is distributed as the sum of $m$ iid geometric random variables with $p = e^{-t}$. Thus

$$\mathbb{P}\left(Y^m(t) > \lambda\right) \leq m \mathbb{P}\left(\text{geom}(e^{-t}) > \frac{\lambda}{m}\right) \leq m \exp\left[-\frac{\lambda}{m} e^{-t}\right].$$

Plugging in $m = \ell_n + \lceil \alpha \rceil, t = t_n^+, \lambda = C n^{\frac{1-\gamma}{2+\alpha}} (\log n)^2$ we get,

$$\omega_n n^\gamma \, \mathbb{P}\left(X_i(t_n^+) \geq C n^{\frac{1-\gamma}{2+\alpha}} (\log n)^2\right) \leq K \omega_n n^\gamma \log n \, n^{-C}$$

which goes to zero for sufficiently large $C$.

■

**Lemma 6.8.19.** *For $C$ large enough as $n \to \infty$,*

$$\mathbb{P}(|\tilde{\mathcal{T}}_n^+(\Delta_n)| \geq \omega_n n^\gamma) \to 0, \qquad \mathbb{P}(M_n(\Delta_n) > \ell_n) \to 0.$$

*Proof.* We first prove the assertion on $|\tilde{\mathcal{T}}_n(\Delta_n)|$. Note the size of the tree grows according to a rate one Yule process. Thus by Lemma 6.5.4, $|\tilde{\mathcal{T}}_n(\Delta_n)| \sim \text{Geom}\left(e^{-\gamma log n - B}\right)$. Thus

$$\mathbb{P}\left(|\tilde{\mathcal{T}}_n^+(\Delta_n)| \geq \omega_n n^\gamma\right) \leq \exp\left[-\omega_n n^\gamma e^{-\gamma \log n - B}\right] \to 0, \qquad \text{as } n \to \infty.$$

For the second assertion, note that for any $0 \leq t \leq \Delta_n$, the rate at which a new vertex is born is $|\tilde{\mathcal{T}}_n^+(t)|$. Since the offspring distribution of each new vertex (before time $\Delta_n$) is a Poisson process, the probability that this new vertex has degree greater than $\ell_n$ conditional on $\tilde{\mathcal{T}}_n^+(t)$ is

$$\mathbb{P}(\text{Poisson}(\Delta_n - t) \geq \ell_n) \leq \mathbb{P}(\text{Poisson}(\Delta_n) \geq \ell_n).$$

Thus writing $N_n(\Delta_n)$ for the number of vertices with degree at least $\ell_n$ by time $\Delta_n$ and recalling that for $t \leq \Delta_n$, $\mathbb{E}(\tilde{\mathcal{T}}_n^+(t)) = e^t$ we have,

$$\mathbb{E}(N_n(\Delta_n)) = \int_0^{\Delta_n} \mathbb{P}(\text{Poisson}(\Delta_n - t) \geq \ell_n) e^t \, dt \leq \frac{e^B}{n^\gamma} \mathbb{P}(\text{Poisson}(\Delta_n) \geq \ell_n).$$

Since $\Delta_n = \gamma \log n + B$ with $\gamma < 1$, exponential tail bounds for the Poisson distribution completes the proof.

■

### 6.9 Proofs: Convergence rates for model without change point

This section is dedicated to proving Theorem 6.3.3 and Theorem 6.3.4.

**Lemma 6.9.1.** *Consider a continuous time branching process with attachment function $f$ that satisfies Assumption 6.2.1. Fix $\beta \in (0, \lambda^*)$. There exist positive constants $C_1$, $C_2$ such that if $h$ solves the renewal equation*

$$h(t) = e^{-\lambda^* t}\phi(t) + \int_0^t h(t-s) e^{-\lambda^* s} \mu_f(ds)$$

*with any $\phi$ satisfying $|\phi(s)| \le C_\phi e^{\beta s}$ for all $s \ge 0$, for some $C > 0$, denoting $h(\infty) = \lim_{t\to\infty} h(t)$, we have for all $t \ge 0$,*

$$|h(\infty) - h(t)| \le C_1 C_\phi e^{-C_2 t}.$$

*Proof.* We will use estimates about quantitative rates of convergence for renewal measures derived in (Bardet et al., 2015) in the setting of the point process with i.i.d. inter-arrival times having distribution $e^{-\lambda^* s}\mu_f(ds)$. By Assumption 6.2.1 (iii), it is clear that the measure $e^{-\lambda^* s}\mu_f(ds)$ satisfies $\int_0^\infty e^{\beta' s}e^{-\lambda^* s}\mu_f(ds) < \infty$ for some $\beta' > 0$ and thus, Assumption 1 of (Bardet et al., 2015) is satisfied. Moreover, for any Borel set $A$ in $[0,1]$, denoting by $E$ the first time the root reproduces (which has an exponential distribution with rate $f(0)$), note that

$$\mu_f(A) \ge \mathbb{E}(\mathbb{1}\{E \in A\}) = \int_A f(0)e^{-f(0)x}dx \ge f(0)e^{-f(0)}\int_A dx$$

and consequently, the distribution of the inter-arrival time is *spread out* in the sense of Assumption 2 of (Bardet et al., 2015) taking $c = 1/2, L = 1/2$ and $\widetilde{\eta} = f(0)e^{-(\lambda^* + f(0))}$. Thus, Corollary 1 of (Bardet et al., 2015) holds for the point process under consideration. For any $x \ge 0$, denote by $U^x$ the renewal measure corresponding to the associated point process with time started at $x$. The stationary version of this point process corresponds to a random starting time whose law is $\mu^*(ds) = m^{\star -1}se^{-\lambda^* s}\mu_f(ds)$ (called the *stationary delay distribution*), where $m^\star = \int_0^\infty ue^{-\lambda^* u}\mu_f(du)$. From translation invariance, it follows that the renewal measure associated to this stationary version is given by $U^*(ds) = m^{\star -1}ds$. By Corollary 1 of (Bardet et al., 2015), there exist constants $C, C' > 0$ and $\beta'' < \beta'$ such that for any Borel set $D \subset (0,\infty)$ and any $x, t \ge 0$,

$$|U^x(D+t) - U^0(D+t)| \le Ce^{\beta'' x}e^{-C't}(U^0((0,\sup D)) + 1).$$

Integration both sides of the above relation over $x$ with respect to the stationary delay distribution $\mu^*(dx)$ and using Fubini's theorem and the fact that $\int_0^\infty e^{\beta' s}e^{-\lambda^* s}\mu_f(ds) < \infty$, we obtain

$$|U^*(D+t) - U^0(D+t)| \le Ce^{-C't}(U^0((0,\sup D)) + 1).$$

This, in turn, implies that for ant $t \geq 0$, if $U^*_{M,t}$ and $U^0_{M,t}$ denote the measures defined by $U^*_{M,t}(D) = U^*(D+t)$ and $U^0_{M,t}(D) = U^0(D+t)$ for any Borel set $D \subset [0,M]$, then using the fact that $\lim_{t\to\infty} t^{-1}U^0([0,t]) = \frac{1}{m^\star}$ (which follows from the elementary renewal theorem),

$$||U^*_{M,t} - U^0_{M,t}||_{TV} \leq CMe^{-C't}. \tag{6.9.1}$$

From standard results in renewal theory, observe that $h(\infty) = \int_0^\infty e^{-\lambda^* s}\phi(s)U^*(ds)$ and $h(t) = \int_0^t e^{-\lambda^*(t-s)}\phi(t-s)U^0(ds)$. Thus, for $t \geq 0$,

$$\begin{aligned} |h(\infty) - h(t)| &= \left| \int_0^\infty e^{-\lambda^* s}\phi(s)U^*(ds) - \int_0^t e^{-\lambda^*(t-s)}\phi(t-s)U^0(ds) \right| \\ &\leq \left| \int_0^t e^{-\lambda^* s}\phi(s)U^*(ds) - \int_0^t e^{-\lambda^*(t-s)}\phi(t-s)U^0(ds) \right| + \int_t^\infty e^{-\lambda^* s}\phi(s)U^*(ds). \end{aligned} \tag{6.9.2}$$

As $|\phi(s)| \leq C_\phi e^{\beta s}$ for all $s$,

$$\int_t^\infty e^{-\lambda^* s}\phi(s)U^*(ds) \leq C_\phi m^{\star-1}\int_t^\infty e^{-(\lambda^*-\beta)s}ds = \frac{C_\phi}{m^\star(\lambda^*-\beta)}e^{-(\lambda^*-\beta)t}. \tag{6.9.3}$$

To estimate the first term in the bound (6.9.2), note that for $t \geq 0$,

$$\begin{aligned} &\left| \int_0^t e^{-\lambda^* s}\phi(s)U^*(ds) - \int_0^t e^{-\lambda^*(t-s)}\phi(t-s)U^0(ds) \right| \\ &\qquad = \left| \int_0^t e^{-\lambda^*(t-s)}\phi(t-s)U^*(ds) - \int_0^t e^{-\lambda^*(t-s)}\phi(t-s)U^0(ds) \right| \\ &\qquad \leq \int_0^{t/2} e^{-\lambda^*(t-s)}\phi(t-s)U^*(ds) + \int_0^{t/2} e^{-\lambda^*(t-s)}\phi(t-s)U^0(ds) \\ &\qquad + \left| \int_{t/2}^t e^{-\lambda^*(t-s)}\phi(t-s)U^*(ds) - \int_{t/2}^t e^{-\lambda^*(t-s)}\phi(t-s)U^0(ds) \right| \\ &\leq C_\phi e^{-(\lambda^*-\beta)t/2}U^*([0,t/2]) + C_\phi e^{-(\lambda^*-\beta)t/2}U^0([0,t/2]) + C_\phi||U^*_{t/2,t/2} - U^0_{t/2,t/2}||_{TV} \leq C_1' C_\phi e^{-C_2' t} \end{aligned} \tag{6.9.4}$$

for constants $C_1', C_2' > 0$ not depending on $\phi$, where we have used (6.9.1) along with the observations that $U^*([0,t/2] = \frac{t}{2m^\star}$ and $\lim_{t\to\infty} t^{-1}U^0([0,t/2]) = \frac{1}{2m^\star}$. The lemma follows by using (6.9.3) and (6.9.4) in (6.9.2).

■

*Proof of Theorem 6.3.4.* In the proof, $C, C', C'', C_1, C_2, \beta', \beta$ will denote generic positive constants not depending on $b_\phi$ and the specific choice of $\phi$. Following (Nerman, 1981), we write $x = (x', i)$ to denote that $x$ is the $i$-th child of $x'$ and define for any $t, c \geq 0$,

$$\mathscr{I}(t) = \{x = (x', i) : \sigma_{x'} \leq t \text{ and } t < \sigma_x < \infty\}, \;\; \mathscr{I}(t, c) = \{x = (x', i) : \sigma_{x'} \leq t \text{ and } t + c < \sigma_x < \infty\}.$$

Let $T_t$ denote the number of vertices born by time $t$ and let $\mathscr{A}_n$ be the filtration generated by the entire life histories of the first $n$ vertices (see (Nerman, 1981) for detailed definitions). Define $\mathscr{F}_t = \mathscr{A}_{T_t}$. For any $s > 0$, write $\phi = \phi_s + \phi'_s$ where $\phi_s(u) = \phi(u)\mathbb{1}\{u < s\}$ and $\phi'_s(u) = \phi(u)\mathbb{1}\{u \geq s\}$. Note that

$$\begin{aligned}\mathbb{E}\left|e^{-\lambda^* t} Z_f^\phi(t) - W_\infty M_f^\phi(\infty)\right| \leq \mathbb{E}\left|e^{-\lambda^* t}\left(Z_f^\phi(t) - Z_f^{\phi_s}(t)\right)\right| + \mathbb{E}\left|e^{-\lambda^* t} Z_f^{\phi_s}(t) - W_\infty M_f^{\phi_s}(\infty)\right| \\ + \mathbb{E}\left(\left|M_f^{\phi_s}(\infty) - M_f^\phi(\infty)\right| W_\infty\right). \quad (6.9.5)\end{aligned}$$

The third term in the bound (6.9.5) can be bounded as

$$\begin{aligned}\mathbb{E}\left(\left|M_f^{\phi_s}(\infty) - M_f^\phi(\infty)\right| W_\infty\right) = M_f^{\phi'_s}(\infty) = \frac{1}{m^\star}\int_s^\infty e^{-\lambda * u}\,\mathbb{E}\left(\phi(u)\right) du \\ \leq \frac{b_\phi}{m^\star}\int_s^\infty e^{-\lambda * u}\,\mathbb{E}\left(\xi_f(u) + 1\right) du \leq C b_\phi e^{-(\lambda^* - \beta')s} \quad (6.9.6)\end{aligned}$$

for some $\beta' < \lambda^*$ by virtue of Assumption 6.2.1 (iii). The first term in the bound (6.9.5) can be bounded as

$$\mathbb{E}\left|e^{-\lambda^* t}\left(Z_f^\phi(t) - Z_f^{\phi_s}(t)\right)\right| \quad = \quad \mathbb{E}\left(e^{-\lambda^* t} Z_f^{\phi'_s}(t)\right) \quad \leq \quad \left|M_f^{\phi'_s}(t) - M_f^{\phi'_s}(\infty)\right| \quad + \quad M_f^{\phi'_s}(\infty). \quad (6.9.7)$$

By the fact that $M_f^{\phi'_s}(t)$ satisfies the renewal equation (6.3.3) (with $\phi'_s$ in place of $\phi$) and Lemma 6.9.1, for $t \geq 0$,

$$\left|M_f^{\phi'_s}(t) - M_f^{\phi'_s}(\infty)\right| \leq C_1 b_\phi e^{-C_2 t}.$$

Using this estimate and (6.9.6) in (6.9.7), we obtain

$$\mathbb{E}\left|e^{-\lambda^* t}\left(Z_f^\phi(t) - Z_f^{\phi_s}(t)\right)\right| \leq C_1 b_\phi e^{-C_2 t} + C b_\phi e^{-(\lambda^* - \beta')s}. \quad (6.9.8)$$

Using (6.9.6) and (6.9.8) in (6.9.5), for any $t, s \geq 0$,

$$\mathbb{E}\left|e^{-\lambda^* t} Z_f^{\phi}(t) - W_\infty M_f^{\phi}(\infty)\right| \leq \mathbb{E}\left|e^{-\lambda^* t} Z_f^{\phi_s}(t) - W_\infty M_f^{\phi_s}(\infty)\right| + C_1 b_\phi e^{-C_2 t} + 2C b_\phi e^{-(\lambda^* - \beta')s}. \tag{6.9.9}$$

Now, we estimate the first term in the above bound. Observe that as $\phi_s(u) = 0$ for all $u \geq s$, every individual that contributes to $Z_f^{\phi_s}(t+s)$ must be born after time $t$. Therefore,

$$Z_f^{\phi_s}(t+s) = \sum_{x \in \mathscr{I}(t)} Z_{f,x}^{\phi_s}(t+s-\sigma_x)$$

where for any vertex $x$ and any $u \geq 0$, $Z_{f,x}^{\phi_s}(u)$ denotes the aggregate $\phi$-score at time $\sigma_x + u$ treating the vertex $x$ as the root.

For $t, c \geq 0$ such that $s \geq c$, write

$$X(t,s,c) = \sum_{x \in \mathscr{I}(t) \setminus \mathscr{I}(t,c)} e^{-\lambda^* \sigma_x} \left(e^{-\lambda^*(t+s-\sigma_x)} Z_{f,x}^{\phi_s}(t+s-\sigma_x) - M_f^{\phi_s}(t+s-\sigma_x)\right).$$

and write $W_t = \sum_{x \in \mathscr{I}(t)} e^{-\lambda^* \sigma_x}$, $W_{t,c} = \sum_{x \in \mathscr{I}(t,c)} e^{-\lambda^* \sigma_x}$. Following equation (3.36) in (Nerman, 1981), we obtain

$$\begin{aligned}\left|e^{-\lambda^*(t+s)} Z_f^{\phi_s}(t+s) - W_\infty M_f^{\phi_s}(\infty)\right| &\leq |X(t,s,c)| + \sum_{x \in \mathscr{I}(t) \setminus \mathscr{I}(t,c)} e^{-\lambda^* \sigma_x} \left|M_f^{\phi_s}(t+s-\sigma_x) - M_f^{\phi_s}(\infty)\right| \\ &\quad + \left|\sum_{x \in \mathscr{I}(t,c)} e^{-\lambda^* \sigma_x} \left(e^{-\lambda^*(t+s-\sigma_x)} Z_{f,x}^{\phi_s}(t+s-\sigma_x) - M_f^{\phi_s}(\infty)\right)\right| + M_f^{\phi_s}(\infty) |W_t - W_\infty|. \end{aligned} \tag{6.9.10}$$

Note that

$$\mathrm{Var}(X(t,s,c)|\mathscr{F}_t) = \sum_{x \in \mathscr{I}(t) \setminus \mathscr{I}(t,c)} e^{-2\lambda^* \sigma_x} V_f^{\phi_s}(t+s-\sigma_x) \tag{6.9.11}$$

where $V_f^{\phi_s}(t) = \mathrm{Var}\left(e^{-\lambda^* t} Z_f^{\phi_s}(t)\right)$. Recall $m_f^{\phi_s}(t) = \mathbb{E}\left(Z_f^{\phi_s}(t)\right)$ and $v_f^{\phi_s}(t) = \mathrm{Var}\left(Z_f^{\phi_s}(t)\right)$. From Theorem 3.2 of (Jagers and Nerman, 1984a), $v_f^{\phi_s}(t) = h \star U(t)$, where

$$h(t) = \mathrm{Var}\left(\phi_s(t) + \int_0^t m_f^{\phi_s}(t-u) \xi_f(du)\right)$$

and $U(\cdot) = \sum_{\ell=0}^{\infty} \mu_f^{\star \ell}(\cdot)$ denotes the renewal measure.

As $\phi_s(t) \le b_\phi(\xi_f(t)+1)$ for all $t$ and Assumption 6.3.2 holds,

$$e^{-2\lambda^* t}\mathbb{E}(\phi_s(t))^2 \le (b_\phi)^2\mathbb{E}\left(e^{-\lambda^* t}(1+\xi_f(t))\right)^2 \le 2(b_\phi)^2\mathbb{E}\left(e^{-2\lambda^* t}+\lambda^{*2}\left(\int_t^\infty e^{-\lambda^* u}\xi_f(u)du\right)^2\right) \le C(b_\phi)^2. \tag{6.9.12}$$

As $\mathbb{E}\left(\xi_f(t)+1\right) \le Ce^{\beta' t}$ by Assumption 6.2.1 (iii), therefore $\mathbb{E}\left(\phi_s(t)\right) \le b_\phi\mathbb{E}\left(\xi_f(t)+1\right) \le b_\phi Ce^{\beta' t}$. Hence, by the fact that $M_f^{\phi_s}(t)$ satisfies the renewal equation (6.3.3) and Lemma 6.9.1, for $t \ge 0$,

$$\left|M_f^{\phi_s}(t) - M_f^{\phi_s}(\infty)\right| \le C_1 b_\phi e^{-C_2 t}. \tag{6.9.13}$$

Moreover,

$$M_f^{\phi_s}(\infty) = \frac{\int_0^\infty e^{-\lambda^* u}\mathbb{E}(\phi_s(u))du}{m^\star} \le \frac{b_\phi\int_0^\infty \mathbb{E}\left(e^{-\lambda^* u}(1+\xi_f(u))\right)du}{m^\star} \le Cb_\phi. \tag{6.9.14}$$

Using (6.9.13) and (6.9.14), we obtain for all $t \ge 0$,

$$M_f^{\phi_s}(t) \le C' b_\phi. \tag{6.9.15}$$

From (6.9.12) and (6.9.15), we conclude for all $t \ge 0$,

$$\begin{aligned} e^{-2\lambda^* t}h(t) &= \mathrm{Var}\left(e^{-\lambda^* t}\phi_s(t) + \int_0^t e^{-\lambda^*(t-u)}m_f^{\phi_s}(t-u)e^{-\lambda^* u}\xi_f(du)\right) \\ &\le 2e^{-2\lambda^* t}\mathbb{E}(\phi_s(t))^2 + 2\mathbb{E}\left(\int_0^t M_f^{\phi_s}(t-u)e^{-\lambda^* u}\xi_f(du)\right)^2 \\ &\le 2C(b_\phi)^2 + 2(Cb_\phi)^2\mathbb{E}\left(\int_0^\infty e^{-\lambda^* u}\xi_f(du)\right)^2 \le C'(b_\phi)^2. \end{aligned}$$

Thus, for all $t \ge 0$,

$$\begin{aligned} V_f^{\phi_s}(t) = \int_0^\infty e^{-2\lambda^*(t-u)}h(t-u)e^{-2\lambda^* u}U(du) &\le C'(b_\phi)^2\int_0^\infty e^{-2\lambda^* u}U(du) = C'(b_\phi)^2\sum_{\ell=0}^\infty \hat{\mu}_f(2\lambda^*)^\ell \\ &= \frac{C'(b_\phi)^2}{1-\hat{\mu}_f(2\lambda^*)} = C''(b_\phi)^2. \end{aligned} \tag{6.9.16}$$

Using this bound in (6.9.11), we obtain

$$\mathbb{E}\left(\mathrm{Var}(X(t,s,c)|\mathscr{F}_t)\right) \le C''(b_\phi)^2\mathbb{E}\left(\sum_{x\in\mathscr{I}(t)\setminus\mathscr{I}(t,c)} e^{-2\lambda^*\sigma_x}\right) \le C''(b_\phi)^2 e^{-\lambda^* t}\mathbb{E}(W_t) = C''(b_\phi)^2 e^{-\lambda^* t}.$$

Moreover, $\mathbb{E}(X(t,s,c)|\mathscr{F}_t) = 0$. Thus, we obtain

$$\mathbb{E}|X(t,s,c)| \le \sqrt{\mathbb{E}(X(t,s,c))^2} = \sqrt{\mathrm{Var}(X(t,s,c))} \le \sqrt{C''} b_\phi e^{-\lambda^* t/2}. \tag{6.9.17}$$

Using (6.9.13),

$$\mathbb{E}\left( \sum_{x \in \mathscr{I}(t) \setminus \mathscr{I}(t,c)} e^{-\lambda^* \sigma_x} \left| M_f^{\phi_s}(t+s-\sigma_x) - M_f^{\phi_s}(\infty) \right| \right) \le C_1 b_\phi e^{-C_2(s-c)} \mathbb{E}(W_t) = C_1 b_\phi e^{-C_2(s-c)}. \tag{6.9.18}$$

To estimate the third term in the bound (6.9.10), observe that upon conditioning on $\mathscr{F}_t$ and noting that $\sup_{t<\infty} M_f^{\phi_s}(t) \le C' b_\phi$,

$$\begin{aligned} &\mathbb{E}\left( \left| \sum_{x \in \mathscr{I}(t,c)} e^{-\lambda^* \sigma_x} \left( e^{-\lambda^*(t+s-\sigma_x)} Z_{f,x}^{\phi_s}(t+s-\sigma_x) - M_f^{\phi_s}(\infty) \right) \right| \right) \\ &\qquad \le \mathbb{E}\left( \sum_{x \in \mathscr{I}(t,c)} e^{-\lambda^* \sigma_x} \left( M_f^{\phi_s}(t+s-\sigma_x) + M_f^{\phi_s}(\infty) \right) \right) \le C' b_\phi \mathbb{E}(W_{t,c}). \end{aligned} \tag{6.9.19}$$

Consider the characteristic $\phi^c(v) = e^{\lambda^* v}\left( \int_{v+c}^\infty e^{-\lambda^* u} \xi_f(du) \right)$, $v \ge 0$. Then $W_{t,c} = e^{-\lambda^* t} Z_f^{\phi^c}(t)$. Note that

$$\begin{aligned} \mathbb{E}(\phi^c(t)) &= e^{\lambda^* t} \mathbb{E}\left( \int_{t+c}^\infty e^{-\lambda^* u} \xi_f(du) \right) = e^{\lambda^* t} \mathbb{E}\left( \int_{t+c}^\infty \lambda^* e^{-\lambda^* v} (\xi_f(v) - \xi_f(t+c)) dv \right) \\ &\le e^{\lambda^* t} \mathbb{E}\left( \int_{t+c}^\infty \lambda^* e^{-\lambda^* v} \xi_f(v) dv \right) \le C e^{\lambda^* t} \left( \int_{t+c}^\infty \lambda^* e^{-\lambda^* v} e^{\beta' v} dv \right) \le \frac{C\lambda^* e^{\lambda^* t}}{\lambda^* - \beta'} e^{-(\lambda^* - \beta')t} = \frac{C\lambda^* e^{\beta' t}}{\lambda^* - \beta'}. \end{aligned}$$

Hence, by Lemma 6.9.1,

$$\left| M_f^{\phi^c}(t) - M_f^{\phi^c}(\infty) \right| \le C_1 e^{-C_2 t}. \tag{6.9.20}$$

Moreover, by Lemma 3.5 of (Nerman, 1981),

$$M_f^{\phi^c}(\infty) = \frac{\int_c^\infty (1 - \mu_{f,\lambda^*}(u)) du}{\int_0^\infty (1 - \mu_{f,\lambda^*}(u)) du}$$

where $\mu_{f,\lambda^*}(u) = \int_0^u e^{-\lambda^* v} \mu_f(dv)$. Now, for any $u \ge 0$,

$$1 - \mu_{f,\lambda^*}(u) = \int_u^\infty e^{-\lambda^* v} \mu_f(dv) \le \int_u^\infty \lambda^* e^{-\lambda^* v} \mu_f(v) dv \le C \int_u^\infty \lambda^* e^{-\lambda^* v} e^{\beta' v} dv = \frac{C\lambda^*}{\lambda^* - \beta'} e^{-(\lambda^* - \beta')u}$$

and hence,

$$\int_c^\infty (1-\mu_{f,\lambda^*}(u))\,du \le \int_c^\infty \frac{C\lambda^*}{\lambda^*-\beta'} e^{-(\lambda^*-\beta')u}\,du = \frac{C\lambda^*}{(\lambda^*-\beta')^2} e^{-(\lambda^*-\beta')c}.$$

This bound implies that there exists $C>0$ such that for all $c>0$,

$$M_f^{\phi^c}(\infty) \le Ce^{-(\lambda^*-\beta')c}. \tag{6.9.21}$$

Combining (6.9.20) and (6.9.21),

$$\mathbb{E}(W_{t,c}) = M_f^{\phi^c}(t) \le C_1 e^{-C_2 t} + Ce^{-(\lambda^*-\beta')c}.$$

Using this in (6.9.19), we get

$$\mathbb{E}\left(\left|\sum_{x\in\mathscr{I}(t,c)} e^{-\lambda^*\sigma_x}\left(e^{-\lambda^*(t+s-\sigma_x)}Z_{f,x}^{\phi_s}(t+s-\sigma_x) - M_f^{\phi_s}(\infty)\right)\right|\right) \le C' b_\phi \left(e^{-C_2 t} + e^{-(\lambda^*-\beta')c}\right). \tag{6.9.22}$$

To estimate the last term in the bound (6.9.10), observe that for any $t\ge 0$, $W_\infty = \sum_{x\in\mathscr{I}(t)} e^{-\lambda^*\sigma_x} W_\infty^x$, where $W_\infty^x$ corresponds to $W_\infty$ treating vertex $x$ as the root (and hence are i.i.d and have the same distribution as $W_\infty$). Moreover, by Theorem 4.1 of (Jagers and Nerman, 1984a), $\mathrm{Var}(W_\infty)<\infty$. Using these observations,

$$\begin{aligned}\mathbb{E}(W_t - W_\infty)^2 = \mathbb{E}\left(\sum_{x\in\mathscr{I}(t)} e^{-\lambda^*\sigma_x}(1-W_\infty^x)\right)^2 &= \mathrm{Var}(W_\infty)\,\mathbb{E}\left(\sum_{x\in\mathscr{I}(t)} e^{-2\lambda^*\sigma_x}\right)\\ &\le \mathrm{Var}(W_\infty)\, e^{-\lambda^* t}\mathbb{E}(W_t) = \mathrm{Var}(W_\infty)\, e^{-\lambda^* t}.\end{aligned}$$

Together with the fact that $\sup_{t<\infty} M_f^{\phi_s}(t) \le C' b_\phi$, this implies that for $t\ge 0$,

$$\mathbb{E}\left|M_f^{\phi_s}(\infty)\,|W_t - W_\infty|\right| \le \sqrt{\mathbb{E}\left(M_f^{\phi_s}(\infty)\,|W_t-W_\infty|\right)^2} \le C' b_\phi e^{-\lambda^* t/2}. \tag{6.9.23}$$

Using (6.9.17), (6.9.18), (6.9.22) and (6.9.23) and the bound (6.9.10), we obtain $D, D_1, D_2, D_3 > 0$ not dependin on $b_\phi, t, s, c$ such that

$$\mathbb{E}\left(\left|e^{-\lambda^*(t+s)}Z_f^{\phi_s}(t+s) - W_\infty M_f^{\phi_s}(\infty)\right|\right) \le D b_\phi \left(e^{-D_1 t} + e^{-D_2 c} + e^{-D_3(s-c)}\right). \tag{6.9.24}$$

Using (6.9.24) in (6.9.9), we obtain

$$\mathbb{E}\left|e^{-\lambda^* t} Z_f^{\phi}(t) - W_\infty M_f^{\phi}(\infty)\right| \le D b_\phi \left(e^{-D_1 t} + e^{-D_2 c} + e^{-D_3(s-c)}\right) + C_1 b_\phi e^{-C_2 t} + 2C b_\phi e^{-(\lambda^* - \beta')s}.$$

The lemma now follows by taking $s = t$ and $c = t/2$.

■

Recall $\lambda_\ell, \lambda_\ell^{(k)}$ for $k, \ell \ge 0$ from (6.3.4), with $f_1$ replaced by $f$ (as this section considers the model without change point).

**Lemma 6.9.2.** *Consider a continuous time branching process with attachment function $f$ that satisfies Assumptions 6.2.1, 6.3.1 and 6.3.2. There exist $\omega_1, \epsilon^* \in (0,1)$ and positive constants $C, \omega_2$ such that for all $\epsilon \le \epsilon^*$ and all $T \in \left[\frac{1-\epsilon}{\lambda^*}\log n, \frac{1+\epsilon}{\lambda^*}\log n\right]$,*

$$\mathbb{E}\left(n^{\omega_1} \sup_{t \in [0, 2\epsilon \log n/\lambda^*]} \left| e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \lambda_\ell(t) D(\ell, T) - \sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell W_\infty \right| \right) \le C n^{-\omega_2}$$

*and for any $k \ge 0$,*

$$\mathbb{E}\left(n^{\omega_1} \sup_{t \in [0, 2\epsilon \log n/\lambda^*]} \left| e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) D(\ell, T) - \sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) p_\ell W_\infty \right| \right) \le C(k+1) n^{-\omega_2}.$$

*Proof.* For any $t$, consider the characteristic $\phi(s) = \sum_{\ell=0}^{\infty} \lambda_\ell(t) \mathbb{1}\{\xi_f(s) = \ell\}$. Then $Z_f^{\phi}(s) = \sum_{\ell=0}^{\infty} \lambda_\ell(t) D(\ell, s)$. As $\phi$ satisfies the hypothesis of Theorem 6.3.4 with $b_\phi = Ce^{\lambda^* t}$ for some $C > 0$ (which is a consequence of $\lim_{t\to\infty} e^{-\lambda^* t}\lambda_\ell(t) = \frac{w_\ell}{\lambda^* m^*}$), for any $\epsilon \in (0,1)$, any $t \in [0, 2\epsilon \log n/\lambda^*]$ and any $T \in \left[\frac{1-\epsilon}{\lambda^*}\log n, \frac{1+\epsilon}{\lambda^*}\log n\right]$,

$$\mathbb{E}\left(\left| e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \lambda_\ell(t) D(\ell, T) - \sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell W_\infty \right|\right) \le C_1 C e^{\lambda^* t} e^{-\frac{C_2(1-\epsilon)}{\lambda^*}\log n} \le C_1 C e^{2\epsilon \log n} e^{-\frac{C_2(1-\epsilon)}{\lambda^*}\log n}.$$

Therefore, choosing $\epsilon^*$ small enough, there exists $\theta_1 > 0$ such that for any $\epsilon \le \epsilon^*$, any $t \in [0, 2\epsilon \log n/\lambda^*]$ and any $T \in \left[\frac{1-\epsilon}{\lambda^*}\log n, \frac{1+\epsilon}{\lambda^*}\log n\right]$,

$$\mathbb{E}\left(\left| e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \lambda_\ell(t) D(\ell, T) - \sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell W_\infty \right|\right) \le \frac{1}{n^{\theta_1}}. \tag{6.9.25}$$

Take any $\theta_2 \in (0,\theta_1)$ and a partition of $[0, 2\epsilon \log n/\lambda^*]$ into $t_0 < t_1 < \cdots < t_{\lfloor (2\epsilon \log n/\lambda^*) n^{\theta_2} \rfloor + 1}$ of mesh $n^{-\theta_2}$. By Lemma 6.7.4, for any $j$ and any $t \in [t_j, t_{j+1}]$, there exist constants $C, C' > 0$ independent of $\epsilon, n$ such that

$$\begin{aligned}
&\left|\left| e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \lambda_\ell(t) D(\ell,T) - \sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell W_\infty \right| - \left| e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \lambda_\ell(t_j) D(\ell,T) - \sum_{\ell=0}^{\infty} \lambda_\ell(t_j) p_\ell W_\infty \right|\right| \\
&\qquad \le e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \left|\lambda_\ell(t) - \lambda_\ell(t_j)\right| D(\ell,T) + \sum_{\ell=0}^{\infty} \left|\lambda_\ell(t) - \lambda_\ell(t_j)\right| p_\ell W_\infty \\
&\qquad \le \frac{C n^{C'\epsilon}}{n^{1-\epsilon+\theta_2}} \sum_{\ell=0}^{\infty} (\ell+1) D(\ell,T) + \frac{C}{n^{\theta_2}} \sum_{\ell=0}^{\infty} (\ell+1) p_\ell W_\infty \\
&\qquad \le \frac{2C}{n^{1-(1+C')\epsilon+\theta_2}} Z(T) + \frac{2C}{n^{\theta_2}} W_\infty. \quad (6.9.26)
\end{aligned}$$

Using (6.9.25), (6.9.26) and the union bound, we obtain for any $\omega' > 0$,

$$\begin{aligned}
&\mathbb{E}\left( n^{\omega'} \sup_{t \in [0, 2\epsilon \log n/\lambda^*]} \left| e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \lambda_\ell(t) D(\ell,T) - \sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell W_\infty \right| \right) \\
&\qquad \le \mathbb{E}\left( n^{\omega'} \sup_{1 \le j \le \lfloor (2\epsilon \log n/\lambda^*) n^{\theta_2} \rfloor + 1} \left| e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \lambda_\ell(t_j) D(\ell,T) - \sum_{\ell=0}^{\infty} \lambda_\ell(t_j) p_\ell W_\infty \right| \right) \\
&\qquad\qquad + \mathbb{E}\left( \frac{2C n^{\omega'}}{n^{1-(1+C')\epsilon+\theta_2}} Z(T) + \frac{2C n^{\omega'}}{n^{\theta_2}} W_\infty \right) \\
&\qquad \le n^{\omega'} \sum_{j=0}^{\lfloor (2\epsilon \log n/\lambda^*) n^{\theta_2} \rfloor + 1} \mathbb{E}\left( \left| e^{-\lambda^* T} \sum_{\ell=0}^{\infty} \lambda_\ell(t_j) D(\ell,T) - \sum_{\ell=0}^{\infty} \lambda_\ell(t_j) p_\ell W_\infty \right| \right) \\
&\qquad\qquad + n^{\omega'} \mathbb{E}\left( \frac{2C}{n^{1-(1+C')\epsilon+\theta_2}} Z(T) + \frac{2C}{n^{\theta_2}} W_\infty \right) \le \frac{C'' \epsilon \log n}{n^{\theta_1-\theta_2-\omega'}} + \frac{C''}{n^{\theta_2-(2+C')\epsilon-\omega'}} + \frac{C''}{n^{\theta_2-\omega'}}
\end{aligned}$$

for some constant $C'' > 0$. Taking $\epsilon^* < \theta_2/(2+C')$ and any $\omega' < \min\{\theta_1 - \theta_2, \theta_2 - (2+C')\epsilon^*, 1\}$, this proves the first assertion in the lemma. The second assertion follows similarly upon noting that $\lambda_\ell^{(k)} \le \lambda_\ell$ for each $k \ge 0$ (and thus the constant $C$ in the expectation bound can be chosen uniformly over $k$) and using Corollary 6.7.6 in place of Lemma 6.7.4 (which accounts for the $(k+1)$ in the bound).

■

*Proof of Theorem 6.3.3.* Take $\epsilon^{**} \le \epsilon^*$ (where $\epsilon^*$ is as in Lemma 6.9.2) and any $\epsilon \le \epsilon^{**}$. We will abbreviate

$$\mathscr{S}_n := \sup_{t\in[0,2\epsilon\log n/\lambda^*]} \left| \sum_{\ell=0}^{\infty} \lambda_\ell(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right) - n^{1-\epsilon} \sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell W_\infty \right|,$$
$$\mathscr{S}_n^{(k)} := \sup_{t\in[0,2\epsilon\log n/\lambda^*]} \left| \sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right) - n^{1-\epsilon} \sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) p_\ell W_\infty \right|.$$

Observe that for any $k \ge 0$, using the fact that $\lambda_\ell(\cdot)$ is an increasing function and $\lambda_\ell(0) = 1$ for each $\ell \ge 0$,

$$\begin{aligned}
&\sup_{t\in[0,2\epsilon\log n/\lambda^*]} \left| \frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)} - \frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) p_\ell}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell} \right| \\
&\qquad \le \frac{\mathscr{S}_n^{(k)}}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)} + \frac{\mathscr{S}_n\left(\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) p_\ell W_\infty\right)}{\left(\sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell W_\infty\right)\left(\sum_{\ell=0}^{\infty} \lambda_\ell(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)\right)} \\
&\qquad \le \frac{\mathscr{S}_n^{(k)}}{\sum_{\ell=0}^{\infty} \lambda_\ell(0) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)} + \frac{\mathscr{S}_n}{\left(\sum_{\ell=0}^{\infty} \lambda_\ell(0) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)\right)} \\
&\qquad = \frac{\mathscr{S}_n^{(k)}}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)} + \frac{\mathscr{S}_n}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)}.
\end{aligned}$$

Recalling $\omega_1$ from Lemma 6.9.2,

$$\begin{aligned}
&n^{\omega_1} \sum_{k=0}^{\infty} 2^{-k} \left( \sup_{t\in[0,2\epsilon\log n/\lambda^*]} \left| \frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)} - \frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) p_\ell}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell} \right| \right) \\
&\qquad \le \frac{n^{1-\epsilon}}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)} \sum_{k=0}^{\infty} 2^{-k} \left( \frac{\mathscr{S}_n^{(k)}}{n^{1-\epsilon-\omega_1}} + \frac{\mathscr{S}_n}{n^{1-\epsilon-\omega_1}} \right).
\end{aligned}$$

Using Lemma 6.9.2, for any $\eta > 0$,

$$\begin{aligned}
\mathbb{P}\left( \sum_{k=0}^{\infty} 2^{-k} \left( \frac{\mathscr{S}_n^{(k)}}{n^{1-\epsilon-\omega_1}} + \frac{\mathscr{S}_n}{n^{1-\epsilon-\omega_1}} \right) > \eta \right) &\le \eta^{-1} \sum_{k=0}^{\infty} 2^{-k} \frac{1}{n^{1-\epsilon-\omega_1}} \mathbb{E}\left( \mathscr{S}_n^{(k)} + \mathscr{S}_n \right) \\
&\le \eta^{-1} \sum_{k=0}^{\infty} 2^{-k}(k+2) C n^{-\omega_2} \le C' \eta^{-1} n^{-\omega_2}
\end{aligned}$$

for positive constants $C, C'$. Moreover, $\frac{n^{1-\epsilon}}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)} \xrightarrow{P} \frac{\lambda^* m^\star}{W_\infty}$ as $n \to \infty$. By Lemma 6.5.8. Combining these observations,

$$n^{\omega_1} \sum_{k=0}^{\infty} 2^{-k} \left( \sup_{t\in[0,2\epsilon\log n/\lambda^*]} \left| \frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)} - \frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) p_\ell}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell} \right| \right) \xrightarrow{P} 0. \tag{6.9.27}$$

Moreover, it is straightforward to check that

$$\begin{aligned}
&\sup_{t\in[0,2\epsilon\log n/\lambda^*]}\left|\frac{D\left(k,\frac{1-\epsilon}{\lambda^*}\log n+t\right)}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n+t\right)}-\frac{\sum_{\ell=0}^{\infty}\lambda_\ell^{(k)}(t)D\left(\ell,\frac{1-\epsilon}{\lambda^*}\log n\right)}{\sum_{\ell=0}^{\infty}\lambda_\ell(t)D\left(\ell,\frac{1-\epsilon}{\lambda^*}\log n\right)}\right| \\
&\qquad\le \frac{1}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)}\sup_{t\in[0,2\epsilon\log n/\lambda^*]}\left|D\left(k,\frac{1-\epsilon}{\lambda^*}\log n+t\right)-\sum_{\ell=0}^{\infty}\lambda_\ell^{(k)}(t)D\left(\ell,\frac{1-\epsilon}{\lambda^*}\log n\right)\right| \\
&\qquad\quad+\frac{1}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)}\sup_{t\in[0,2\epsilon\log n/\lambda^*]}\left|Z\left(\frac{1-\epsilon}{\lambda^*}\log n+t\right)-\sum_{\ell=0}^{\infty}\lambda_\ell(t)D\left(\ell,\frac{1-\epsilon}{\lambda^*}\log n\right)\right|. \quad (6.9.28)
\end{aligned}$$

Abbreviate

$$\begin{aligned}
\hat{\mathscr{S}}_n^{(k)} &:= \sup_{t\in[0,2\epsilon\log n/\lambda^*]}\left|D\left(k,\frac{1-\epsilon}{\lambda^*}\log n+t\right)-\sum_{\ell=0}^{\infty}\lambda_\ell^{(k)}(t)D\left(\ell,\frac{1-\epsilon}{\lambda^*}\log n\right)\right|, \\
\hat{\mathscr{S}}_n &:= \sup_{t\in[0,2\epsilon\log n/\lambda^*]}\left|Z\left(\frac{1-\epsilon}{\lambda^*}\log n+t\right)-\sum_{\ell=0}^{\infty}\lambda_\ell(t)D\left(\ell,\frac{1-\epsilon}{\lambda^*}\log n\right)\right|.
\end{aligned}$$

By conditioning on $\mathscr{F}_n\left(\frac{1-\epsilon}{\lambda^*}\log n\right)$ and applying Lemma 6.7.11, we obtain $\omega_1'\in(0,1),\omega_2'>0$ not depending on $\epsilon$ such that for any $\eta>0$,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=0}^{\infty}2^{-k}\left(\frac{\hat{\mathscr{S}}_n^{(k)}}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)^{1-\omega_1'}}\right)>\eta\;\middle|\;\mathscr{F}_n\left(\frac{1-\epsilon}{\lambda^*}\log n\right)\right) \\
&\qquad=\mathbb{P}\left(\sum_{k=0}^{\infty}2^{-k}\left(\frac{\hat{\mathscr{S}}_n^{(k)}}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)^{1-\omega_1'}}\right)>\sum_{k=0}^{\infty}\left(\frac{3}{2}\right)^{-k}\frac{\eta}{3}\;\middle|\;\mathscr{F}_n\left(\frac{1-\epsilon}{\lambda^*}\log n\right)\right) \\
&\qquad\le\sum_{k=0}^{\infty}\mathbb{P}\left(\frac{\hat{\mathscr{S}}_n^{(k)}}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)^{1-\omega_1'}}>\left(\frac{4}{3}\right)^{k}\frac{\eta}{3}\;\middle|\;\mathscr{F}_n\left(\frac{1-\epsilon}{\lambda^*}\log n\right)\right) \\
&\le Ce^{C'2\epsilon\log n/\lambda^*}\eta^{-2}Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)^{-\omega_2'}\sum_{k=0}^{\infty}(k+1)^2\left(\frac{3}{4}\right)^{2k}=C'n^{2C'\epsilon/\lambda^*}\eta^{-2}Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)^{-\omega_2'} \quad (6.9.29)
\end{aligned}$$

for positive constants $C,C'$. As $\frac{n^{1-\epsilon}}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)}\xrightarrow{P}\frac{\lambda^* m^\star}{W_\infty}$, the bound above converges to zero almost surely if $\epsilon^{**}$ is chosen sufficiently small and $\epsilon\le\epsilon^{**}$. Similarly,

$$\mathbb{P}\left(\sum_{k=0}^{\infty}2^{-k}\left(\frac{\hat{\mathscr{S}}_n}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)^{1-\omega_1'}}\right)>\epsilon\;\middle|\;\mathscr{F}_n\left(\frac{1-\epsilon}{\lambda^*}\log n\right)\right)\quad\le\quad C'n^{2C'\epsilon/\lambda^*}\epsilon^{-2}Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)^{-\omega_2}. \quad (6.9.30)$$

Using (6.9.28), (6.9.29), (6.9.30) and recalling that $\frac{n^{1-\epsilon}}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n\right)} \xrightarrow{P} \frac{\lambda^* m^\star}{W_\infty}$ as $n \to \infty$, we conclude

$$n^{(1-\epsilon)\omega_1'} \sum_{k=0}^{\infty} 2^{-k} \left( \sup_{t\in[0,2\epsilon\log n/\lambda^*]} \left| \frac{D\left(k, \frac{1-\epsilon}{\lambda^*}\log n + t\right)}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n + t\right)} - \frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) D\left(\ell, \frac{1-\epsilon}{\lambda^*}\log n\right)} \right| \right) \xrightarrow{P} 0. \tag{6.9.31}$$

Choosing $\omega^* = \min\{\omega_1, (1-\epsilon)\omega_1'\}$, we conclude from (6.9.27) and (6.9.31) that

$$n^{\omega^*} \sum_{k=0}^{\infty} 2^{-k} \left( \sup_{t\in[0,2\epsilon\log n/\lambda^*]} \left| \frac{D\left(k, \frac{1-\epsilon}{\lambda^*}\log n + t\right)}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n + t\right)} - \frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) p_\ell}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell} \right| \right) \xrightarrow{P} 0. \tag{6.9.32}$$

Finally, we claim that for each $k \geq 0,\ t \geq 0$,

$$\frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) p_\ell}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell} = p_k. \tag{6.9.33}$$

To see this, observe that the following limits hold as $n \to \infty$:

$$\frac{Z\left(\frac{1-\epsilon}{\lambda^*}\log n + t\right)}{n^{1-\epsilon}} \xrightarrow{P} \frac{e^{\lambda^* t} W_\infty}{\lambda^* m^\star}, \quad \frac{D(k, \frac{1-\epsilon}{\lambda^*}\log n + t)}{n^{1-\epsilon}} \xrightarrow{P} \frac{p_k e^{\lambda^* t} W_\infty}{\lambda^* m^\star}$$

and thus,

$$\frac{D\left(k, \frac{1-\epsilon}{\lambda^*}\log n + t\right)}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n + t\right)} \xrightarrow{P} p_k.$$

But from (6.9.32),

$$\frac{D\left(k, \frac{1-\epsilon}{\lambda^*}\log n + t\right)}{Z\left(\frac{1-\epsilon}{\lambda^*}\log n + t\right)} \xrightarrow{P} \frac{\sum_{\ell=0}^{\infty} \lambda_\ell^{(k)}(t) p_\ell}{\sum_{\ell=0}^{\infty} \lambda_\ell(t) p_\ell}.$$

(6.9.33) follows from the above two observations. The lemma now follows from (6.9.32) and (6.9.33).

■

### 6.10 Proofs: Change point detection

Recall $\lambda_\ell, \lambda_\ell^{(k)}$ for $k, \ell \geq 0$ defined in (6.3.4) and the functional $\Phi_a : \mathscr{P} \to \mathscr{P}$ defined for each $a > 0$ in (6.3.5).

**Lemma 6.10.1.** $\lim_{a\to\infty} \Phi_a(\mathbf{p}) = \mathbf{p}^1$ *(where the limit is taken in the coordinate-wise sense).*

*Proof.* For each $k \geq 0$, by Lemma 6.5.8 (ii), $\lim_{t\to\infty} e^{-\lambda_1^* t} m_{f_1}(t) = \frac{1}{\lambda_1^* m^\star}$ and $\lim_{t\to\infty} e^{-\lambda_1^* t} m_{f_1}^{(k)}(t) = \frac{p_k^1}{\lambda_1^* m^\star}$ and consequently,

$$\lim_{t\to\infty} e^{-\lambda_1^* t}\lambda_\ell(t) = \frac{w_\ell}{\lambda_1^* m^\star}, \quad \lim_{t\to\infty} e^{-\lambda_1^* t}\lambda_\ell^{(k)}(t) = \frac{p_k^1 w_\ell}{\lambda_1^* m^\star}. \tag{6.10.1}$$

Moreover, it is easy to see from (6.3.4) that for any $\ell, k \geq 0$, $e^{-\lambda_1^* t}\lambda_\ell(t) \leq \left(\sup_{u\geq 0} e^{-\lambda_1^* u} m_{f_1}(u)\right) w_\ell$ and $e^{-\lambda_1^* t}\lambda_\ell^{(k)}(t) \leq \left(\sup_{u\geq 0} e^{-\lambda_1^* u} m_{f_1}(u)\right) w_\ell$ for all $t \geq 0$ and this bound is finite. By this observation, we can apply the dominated convergence theorem and (6.10.1) in the formula of $\Phi_a(\mathbf{p})$ to obtain the lemma.

■

**Lemma 6.10.2.** *For any $s, t \geq 0$ and any $j, k \geq 0$,*

$$\sum_{\ell=0}^{\infty} \lambda_j^{(\ell)}(t)\lambda_\ell(s) = \lambda_j(s+t), \quad \sum_{\ell=0}^{\infty} \lambda_j^{(\ell)}(t)\lambda_\ell^{(k)}(s) = \lambda_j^{(k)}(s+t).$$

*Consequently, for any $\mathbf{p} \in \mathscr{P}$,*

$$\Phi_s(\Phi_t(\mathbf{p})) = \Phi_{s+t}(\mathbf{p}).$$

*Proof.* We will only prove the first assertion. The second one follows similarly. Denote by $\mathrm{PA}^{(j)}(\cdot)$ the continuous time branching process with attachment function $i \mapsto f_1(i+j)$ and denote by $D_n^{(j)}(\ell, t)$ the number of vertices of degree $\ell$ at time $t$ (excluding the root). Then

$$\mathbb{E}\left(\mathrm{PA}^{(j)}(t+s) \mid \mathscr{F}_n(t)\right) = \sum_{\ell=j}^{\infty} \mathbb{1}\left\{\xi_{f_1}^{(j)}(t) = \ell - j\right\}\left(1 + \int_0^s m_{f_1}(s-v)\mu_{f_1}^{(\ell)}(dv)\right) + \sum_{\ell=0}^{\infty} D_n^{(j)}(\ell, t)\left(1 + \int_0^s m_{f_1}(s-v)\mu_{f_1}^{(\ell)}(dv)\right)$$

where the first term denotes the expected number of vertices born to the root in the process in the time interval $[t, t+s]$ and the second term denotes the expected number of vertices born in the time interval $[t, t+s]$ to those vertices born in the time interval $(0, t]$. Taking expectation on both sides of the above expression and noting that $\lambda_j(t+s) = \mathbb{E}\left(\mathrm{PA}^{(j)}(t+s)\right)$ and $\mathbb{E}\left(D_n^{(j)}(\ell, t)\right) = \int_0^t m_{f_1}^{(\ell)}(t-u)\mu_{f_1}^{(j)}(du)$, we obtain

$$\lambda_j(t+s) = \sum_{\ell=0}^{\infty}\left(\mathbb{P}\left(\xi_{f_1}^{(j)}(t) = \ell - j\right) + \int_0^t m_{f_1}^{(\ell)}(t-u)\mu_{f_1}^{(j)}(du)\right)\left(1 + \int_0^s m_{f_1}(s-v)\mu_{f_1}^{(\ell)}(dv)\right) = \sum_{\ell=0}^{\infty} \lambda_j^{(\ell)}(t)\lambda_\ell(s).$$

To prove the semigroup property, note that for each $k \geq 0$,

$$\left(\Phi_s(\Phi_t(\mathbf{p}))\right)_k = \left(\frac{\sum_{\ell=0}^{\infty}\left(\Phi_t(\mathbf{p})\right)_\ell \lambda_\ell^{(k)}(s)}{\sum_{\ell=0}^{\infty}\left(\Phi_t(\mathbf{p})\right)_\ell \lambda_\ell(s)}\right) = \left(\frac{\sum_{\ell=0}^{\infty}\left(\sum_{j=0}^{\infty} p_j \lambda_j^{(\ell)}(t)\right)\lambda_\ell^{(k)}(s)}{\sum_{\ell=0}^{\infty}\left(\sum_{j=0}^{\infty} p_j \lambda_j^{(\ell)}(t)\right)\lambda_\ell(s)}\right)$$
$$= \frac{\sum_{j=0}^{\infty} p_j \left(\sum_{\ell=0}^{\infty} \lambda_j^{(\ell)}(t)\lambda_\ell^{(k)}(s)\right)}{\sum_{j=0}^{\infty} p_j \left(\sum_{\ell=0}^{\infty} \lambda_j^{(\ell)}(t)\lambda_\ell(s)\right)} = \frac{\sum_{j=0}^{\infty} p_j \lambda_j^{(k)}(s+t)}{\sum_{j=0}^{\infty} p_j \lambda_j(s+t)} = \left(\Phi_{s+t}(\mathbf{p})\right)_k.$$

■

**Lemma 6.10.3.** *For any $a > 0$ and any $\mathbf{p} \in \mathscr{P}$ such that $\mathbf{p} \neq \mathbf{p}^1$, we have $\Phi_a(\mathbf{p}) \neq \mathbf{p}$.*

*Proof.* Suppose there exists $a > 0$ and $\mathbf{p} \neq \mathbf{p}_1$ such that $\Phi_a(\mathbf{p}) = \mathbf{p}$. Then by Lemma 6.10.2, for any $n \geq 1$, $\Phi_{na}(\mathbf{p}) = \mathbf{p}$. Letting $n \to \infty$ and using Lemma 6.10.1, we obtain $\mathbf{p}^1 = \mathbf{p}$ which gives a contradiction.

■

Now we are ready to prove Theorem 6.3.17.

*Proof of Theorem 6.3.17.* Recall $\omega^*$, $\epsilon^{**}$ from Theorem 6.3.3 applied to the branching process with attachment function $f_0$ and fix any $\epsilon \leq \epsilon^{**}$. Let $\lambda_0^*$ denote the associated Malthusian rate. Take any $n_0 \geq 1$ such that $h_n \geq 1/\gamma$ for all $n \geq n_0$. Observe that for any $\eta > 0$ and any $n \geq n_0$,

$$\mathbb{P}\left(n^{\omega^*}\sum_{k=0}^{\infty} 2^{-k} \sup_{1/h_n \leq t \leq \gamma}\left|\frac{D(k, T_{\lfloor nt \rfloor})}{\lfloor nt \rfloor} - p_k^0\right| > \eta\right)$$
$$\leq \mathbb{P}\left(n^{\omega^*}\sum_{k=0}^{\infty} 2^{-k}\left(\sup_{t \in [0, 2\epsilon \log n/\lambda_0^*]}\left|\frac{D\left(\ell, \frac{1-\epsilon}{\lambda_0^*}\log n + t\right)}{Z\left(\frac{1-\epsilon}{\lambda_0^*}\log n + t\right)} - p_k^0\right|\right) > \eta\right)$$
$$+ \mathbb{P}\left(T_{\lfloor n/h_n \rfloor} < \frac{1-\epsilon}{\lambda_0^*}\log n\right) + \mathbb{P}\left(T_{\lfloor n\gamma \rfloor} > \frac{1+\epsilon}{\lambda_0^*}\log n\right).$$

The first term in the above bound converges to zero by Theorem 6.3.3. Further,

$$\mathbb{P}\left(T_{\lfloor n/h_n \rfloor} < \frac{1-\epsilon}{\lambda_0^*}\log n\right) \to 0 \tag{6.10.2}$$

because $\frac{T_{\lfloor n/h_n \rfloor}}{\frac{1}{\lambda_0^*}\log(n/h_n)} \xrightarrow{\mathrm{P}} 1$ as $n \to \infty$ by Lemma 6.5.8 (ii) and by assumption, $\frac{\log h_n}{\log n} \to 0$. Similarly,

$$\mathbb{P}\left(T_{\lfloor n\gamma \rfloor} > \frac{1+\epsilon}{\lambda_0^*}\log n\right) \to 0 \tag{6.10.3}$$

because $\frac{T_{\lfloor n\gamma \rfloor}}{\frac{1}{\lambda_0^*}\log(n\gamma)} \xrightarrow{\mathrm{P}} 1$ as $n \to \infty$. Thus, we conclude

$$n^{\omega^*} \sum_{k=0}^{\infty} 2^{-k} \sup_{1/h_n \le t \le \gamma} \left| \frac{D(k, T_{\lfloor nt \rfloor})}{\lfloor nt \rfloor} - p_k^0 \right| \xrightarrow{\mathrm{P}} 0 \tag{6.10.4}$$

as $n \to \infty$ which, along with the fact that $\omega^* \in (0,1)$, implies

$$n^{\omega^*} \sum_{k=0}^{\infty} 2^{-k} \sup_{1/h_n \le t \le \gamma} \left| \frac{D(k, T_{\lfloor nt \rfloor})}{nt} - \frac{D(k, T_{\lfloor n/h_n \rfloor})}{n/h_n} \right| \xrightarrow{\mathrm{P}} 0.$$

As $\frac{\log b_n}{\log n} \to 0$ as $n \to \infty$, the above implies

$$b_n \sum_{k=0}^{\infty} 2^{-k} \sup_{1/h_n \le t \le \gamma} \left| \frac{D(k, T_{\lfloor nt \rfloor})}{nt} - \frac{D(k, T_{\lfloor n/h_n \rfloor})}{n/h_n} \right| \xrightarrow{\mathrm{P}} 0.$$

From this observation and the definition of $\hat{T}_n$, we conclude that

$$\mathbb{P}\left(\hat{T}_n \ge \gamma\right) \to 1 \text{ as } n \to \infty. \tag{6.10.5}$$

Moreover, by Theorem 6.3.6, for any $t > \gamma$ and any $k \ge 0$, $\left| \frac{D(k, T_{\lfloor tn \rfloor})}{tn} - \left(\Phi_{a_t}(\mathbf{p^0})\right)_k \right| \xrightarrow{\mathrm{P}} 0$ and hence, by (6.10.4) and the dominated convergence theorem, as $n \to \infty$,

$$\sum_{k=0}^{\infty} 2^{-k} \left| \frac{D(k, T_{\lfloor nt \rfloor})}{nt} - \frac{D(k, T_{\lfloor n/h_n \rfloor})}{n/h_n} \right| \xrightarrow{\mathrm{P}} \sum_{k=0}^{\infty} 2^{-k} \left| \left(\Phi_{a_t}(\mathbf{p^0})\right)_k - p_k^0 \right|.$$

As $a_t > 0$ for each $t > \gamma$ and $\mathbf{p}^0 \neq \mathbf{p}^1$, by Lemma 6.10.3, $\Phi_{a_t}(\mathbf{p^0}) \neq \mathbf{p^0}$ and hence, the limit above is strictly positive. From the definition of $\hat{T}_n$ and the above, we conclude that for each $t > \gamma$,

$$\mathbb{P}\left(\hat{T}_n \le t\right) \to 1 \text{ as } n \to \infty. \tag{6.10.6}$$

The theorem follows from (6.10.5) and (6.10.6). ■

CHAPTER 7

# An exposition of the false confidence theorem

A recent paper presents the "false confidence theorem" (FCT) which has potentially broad implications for statistical inference using Bayesian posterior uncertainty. This theorem says that with arbitrarily large (sampling/frequentist) probability, there exists a set which does *not* contain the true parameter value, but which has arbitrarily large posterior probability. Since the use of Bayesian methods has become increasingly popular in applications of science, engineering, and business, it is critically important to understand when Bayesian procedures lead to problematic statistical inferences or interpretations. In this chapter, we consider a number of examples demonstrating the paradoxical nature of false confidence to begin to understand the contexts in which the FCT does (and does not) play a meaningful role in statistical inference. Our examples illustrate that models involving marginalization to non-linear, not one-to-one functions of multiple parameters play a key role in more extreme manifestations of false confidence.

## 7.1 Introduction

In a recent paper, (Balch et al., 2017) presents the phenomenon of "false confidence" associated with Bayesian posterior uncertainty. The authors come about the concept of false confidence from an alarming application to satellite collision risk analysis when estimating the posterior probability of the event that two satellites will collide. They found that increased measurement error of satellite trajectory data leads to decreased posterior probability of satellites colliding. Essentially, as more noise is introduced into trajectory measurements we become less certain about satellite trajectories, and thus the probability of two satellites colliding decreases. However, since a posterior probability is an additive belief function (probabilities of mutually exclusive and collectively exhaustive sets sum to one) the probability of the two satellites not colliding must increase accordingly, making their respective trajectories appear safer. When taken to the extreme, a large enough measurement error will cause an analyst to be (mistak-

enly) certain the satellites will not collide. Conversely, when viewed from a likelihood-based sampling distribution framework, more noise in the trajectory data suggests that the satellite trajectories are less certain and therefore are less likely to collide because of the infinitely large number of possible paths they could each take. This alternative interpretation is not problematic.

More on the specifics and importance of satellite collision risk analysis are provided in (Balch et al., 2017). To study the mechanics behind what is happening at a more fundamental level the authors present what they term the "false confidence theorem" (FCT). This theorem says that with arbitrarily large (sampling/frequentist) probability, there exists a set which does *not* contain the true parameter value, but which has arbitrarily large posterior probability. Such a phenomenon is unsettling for a practitioner making inference based on a posterior distribution. Moreover, the authors prove that false confidence effects all types of epistemic uncertainty represented by additive probability measures. This includes Bayesian posterior probabilities, fiducial probabilities, and probabilities derived from most confidence distributions (Balch et al., 2017).

Our goal is to illustrate the intuition and mechanics of the FCT in simple examples so that we can begin to understand more complicated manifestations of the FCT. Such insight provides a particularly useful contribution to the literature as the use of Bayesian methods becomes more popular. Our contributions in this chapter are the following.

First, we present a simple example to illustrate the mechanics of the FCT with the statistical problem of estimating the support parameter of the $U(0,\theta)$ distribution. This is an example in which the mathematics for the FCT can be worked out analytically and demonstrates where each piece in the statement of the FCT originates from. In most other situations the mathematics cannot be worked out analytically due to the fact that the typical posterior distribution function does not have a readily understood sampling distribution. In the Appendix we provide similar results for a one parameter Gaussian model.

Next, we show that the FCT manifests in an even more pronounced way by extending the first example to a two parameter model, i.e., $U(0,\theta_x)$ and $U(0,\theta_y)$ with $\theta_x \neq \theta_y$, and considering the marginal posterior distribution of the parameter $\psi = \theta_x\theta_y$. This example alludes to the intuition that false confidence is likely at play in situations in which the Gleser-Hwang theorem applies (Gleser and Hwang, 1987). Such examples are characterized in the frequentist paradigm by exhibiting infinitely large confidence intervals required to obtain less than 100 percent coverage (Berger et al., 1999; Gleser and Hwang, 1987). One such famous problem appears in Fieller's theorem (Fieller, 1954) which has been discussed

as recently as the last two meetings of the *Bayesian, Fiducial, and Frequentist Conference* (2017, 2018), and in the forthcoming paper (Fraser et al., 2018).

Finally, we demonstrate that the manifestation of the FCT is immediately apparent in a problem related to Fieller's theorem. We show that in reasonable situations the FCT applies to sets which would be concerning in practice. The contribution of such a striking example of false confidence is worrisome in an era in which Bernstein-von Mises type results are unhesitatingly appealed to even when it may not be appropriate (e.g., certain small sample situations). Such a phenomenon should be properly understood for the appropriate use of Bayesian methodology in practice.

Broadly, the axioms of probability laid down by (Kolmogoroff, 1933) have enabled a rich mathematical theory, however, their suitability for modeling epistemic uncertainty has been met with some discontent, particularly the axiom of additivity (Shafer, 2008). The issue with additivity is that it does not leave room for ignorance (i.e., events are either *true* or *false*) which is a major underpinning of the FCT. Theories of inference which weaken additivity assumptions include inferential models (Martin and Liu, 2016) and imprecise probabilities (Weichselberger, 2000; Gong and Meng, 2017).

The paper is organized as follows. Section 7.2 presents and describes the FCT as given in (Balch et al., 2017). Sections 7.3, 7.4, and 7.5 present and analyze the illustrative examples, and additional analysis is provided in the Appendix. The *R* code to reproduce the numerical results presented in this chapter is provided at `https://github.com/idc9/FalseConfidence`.

### 7.2 Main ideas

This section presents the false confidence theorem from (Balch et al., 2017).

**Theorem 7.2.1** ((Balch et al., 2017))**.** *Consider a countably additive belief function* $\text{Bel}_{\Theta|X}$ *characterized by an epistemic probability density function* $\pi_x(\cdot)$ *on* $\Omega_\theta$ *(the parameter space), with respect to the Lebesgue measure, satisfying* $\sup_{\theta\in\Omega_\theta}\pi_{\mathbf{x}}(\theta)<\infty$*, for* $P_{\mathbf{X}|\theta}$*-almost all* $x$*. Then, for any* $\theta\in\Omega_\theta$*, any* $\alpha\in(0,1)$*, and any* $p\in(0,1)$*, there exists a set* $A\subseteq\Omega_\theta$ *with positive Lebesgue measure such that* $A\not\ni\theta$*, and*

$$P_{\mathbf{X}|\theta}\big(\{X\colon \text{Bel}_{\Theta|X}(A)\geq 1-\alpha\}\big)\geq p. \tag{7.2.1}$$

While Theorem 7.2.1 pertains to any form of epistemic probability, for concreteness we will focus on Bayesian posterior probability. This amounts to considering situations in which

$$\begin{aligned}\text{Bel}_{\Theta|X}(A) &= \int_A \pi_x(\theta)\, d\theta \\ &= \int_A \frac{f_{X|\theta}(X)\pi(\theta)}{\int_{\Omega_\theta} f_{X|\vartheta}(X)\pi(\vartheta)\,\mathrm{d}\vartheta}\, d\theta =: P_{\Theta|X}(A).\end{aligned}$$

To better understand the statement of (7.2.1), Figure 7.2.1 demonstrates the pieces at play. The green region represents an example of a particular $A^c$ as described in the theorem, and each curve represents a particular realization of the posterior distribution (associated with $P_{\Theta|X}$) over the sampling distribution of the data (associated with $P_{\mathbf{X}|\theta}$).

Heuristically speaking, false confidence says that for some set, say $A \subseteq \Omega_\theta$, which does $not$ contain the true parameter value, the (epistemic) posterior probability $P_{\Theta|X}(A)$ can be made arbitrarily large with arbitrarily large (aleatory) sampling/frequentist probability, i.e., with respect to $P_{X|\theta}$. Although the simple existence of such sets $A$ does not immediately raise concerns about statistical inference, for a given situation there may exist practically important sets, such as in the satellite collision risk analysis example of (Balch et al., 2017). Note that these sets $A$ may be particularly concerning for finite sample sizes.

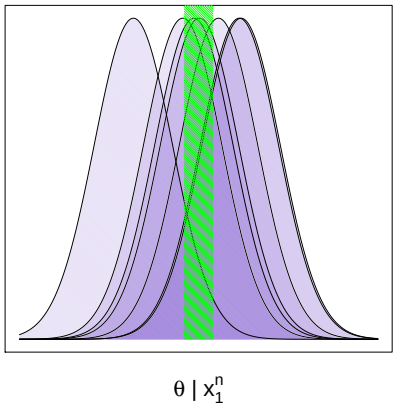



**Figure 7.2.1:** A sample of realizations from the sampling distribution of the posterior density of the mean, $\theta$, for Gaussian data with known variance and normal prior on $\theta$. The green shaded region ($A^c$) is an $\varepsilon$-ball around the true parameter value of $\theta$.

The proof given in (Balch et al., 2017) of the false confidence theorem relies on constructing a neighborhood around the true parameter value. Accordingly, we investigate further the properties of such sets which satisfy Theorem 7.2.1 in a few simple and illustrative examples.

### 7.3 Uniform with Jeffreys' prior

Here we investigate the FCT for uniformly distributed data where the goal is to estimate the support of the distribution. The motivation for considering this example is that it is simple enough that all of the mathematics can be worked out analytically. Let $X_1, \ldots, X_n$ be a random sample from the $\text{U}(0,\theta)$

distribution where $\theta$ is an unknown parameter. Using the Jeffreys' prior, $\pi(\theta) = 1/\theta$, the posterior will be $\theta \mid X_1^n \sim \text{Pareto}(n, X_{(n)})$ where $X_{(n)}$ is the maximum of the observed data (see (Robert, 2007)).

Suppose the true value of $\theta$ is $\theta_0$ and fix $\alpha, p \in (0,1)$. Then by the proof of Theorem 7.2.1 (see (Balch et al., 2017)) there exists $\varepsilon > 0$ such that

$$P_{X_1^n|\theta_0}\left(\{X_1^n \colon P_{\theta|X_1^n}(A_\varepsilon) \geq 1-\alpha\}\right) \geq p, \tag{7.3.1}$$

where $A_\varepsilon := [\theta_0 - \varepsilon, \theta_0 + \varepsilon]^c$, $P_{\theta|X_1^n}$ is the posterior law of $\theta$ (the additive belief function), and $P_{X_1^n|\theta_0}$ is the probability measure associated with the sampling distribution of the data. Note that in this example the Jefferys' prior is a probability matching prior in the Welch-Peers sense (see (Reid et al., 2003)); in particular, the interval $C_x := (-\infty, X^{(n)}\alpha^{-\frac{1}{n}})$ is such that $P_{\theta|X_1^n}(C_x) = 1-\alpha = P_{X_1^n|\theta_0}(X^{(n)}\alpha^{-\frac{1}{n}} \geq \theta_0)$. Since the probability matching prior property in one-dimensions pertains to intervals, this fact provides further justification for considering the Jeffreys' prior for analyzing sets of the form $A_\varepsilon$.

To compute the left side of (7.3.1), first re-express as

$$\begin{aligned}
&P_{X_1^n|\theta_0}\left(F_{\theta|X_1^n}(\theta_0+\varepsilon) - F_{\theta|X_1^n}(\theta_0-\varepsilon) \leq \alpha\right) \\
&\quad = P_{X_1^n|\theta_0}\left(1 - \left(\frac{X_{(n)}}{\theta_0+\varepsilon}\right)^n - \left[1 - \left(\frac{X_{(n)}}{\theta_0-\varepsilon}\right)^n\right]\mathbf{1}\{X_{(n)} \leq \theta_0 - \varepsilon\} \leq \alpha\right) \\
&\quad = P_{X_1^n|\theta_0}\left(\left(\frac{X_{(n)}}{\theta_0-\varepsilon}\right)^n - \left(\frac{X_{(n)}}{\theta_0+\varepsilon}\right)^n \leq \alpha\right) \cdot P_{X_1^n|\theta_0}(X_{(n)} \leq \theta_0 - \varepsilon) \\
&\qquad\qquad + P_{X_1^n|\theta_0}\left(1 - \left(\frac{X_{(n)}}{\theta_0+\varepsilon}\right)^n \leq \alpha\right) \cdot P_{X_1^n|\theta_0}(X_{(n)} > \theta_0 - \varepsilon) \\
&\quad = P_{X_1^n|\theta_0}\left(X_{(n)} \leq \alpha^{\frac{1}{n}}\left(\frac{1}{(\theta_0-\varepsilon)^n} - \frac{1}{(\theta_0+\varepsilon)^n}\right)^{-\frac{1}{n}}\right) \cdot \left(\frac{\theta_0-\varepsilon}{\theta_0}\right)^n \\
&\qquad\qquad + P_{X_1^n|\theta_0}\left(X_{(n)} \geq (1-\alpha)^{\frac{1}{n}}(\theta_0+\varepsilon)\right) \cdot \left[1 - \left(\frac{\theta_0-\varepsilon}{\theta_0}\right)^n\right].
\end{aligned}$$

The second equality comes from the fact that the CDF of the Pareto$(k,m)$ distribution is given by $F(x) = \left(1 - \left(\frac{m}{x}\right)^k\right)\mathbf{1}\{x \geq m\}$. The third equality comes from considering the two cases of the indicator function, and the final equality comes from solving for $X_{(n)}$.

Observe that $\frac{X_{(n)}}{\theta_0} \sim \text{Beta}(n,1)$ (i.e., maximum order statistic of a U$(0,1)$ random sample) which gives $P(X_{(n)} \leq x) = \left(\frac{x}{\theta_0}\right)^n$. Accordingly,

$$\begin{aligned}
&P_{X_1^n|\theta_0}\left(\{X_1^n \colon P_{\theta|X_1^n}([\theta_0-\varepsilon, \theta_0+\varepsilon]) \leq \alpha\}\right) \\
&\quad = \min\left\{1, \alpha\left[\left(\frac{\theta_0}{\theta_0-\varepsilon}\right)^n - \left(\frac{\theta_0}{\theta_0+\varepsilon}\right)^n\right]^{-1}\right\} \cdot \left(\frac{\theta_0-\varepsilon}{\theta_0}\right)^n \\
&\qquad + \left(1 - (1-\alpha)\left(\frac{\theta_0+\varepsilon}{\theta_0}\right)^n\right)\mathbf{1}\left\{\varepsilon \leq \theta_0\left((1-\alpha)^{-\frac{1}{n}} - 1\right)\right\} \cdot \left[1 - \left(\frac{\theta_0-\varepsilon}{\theta_0}\right)^n\right].
\end{aligned} \tag{7.3.2}$$

Setting the right side of equation (7.3.2) equal to $p$ gives $p$ as a function of the $\alpha$, $n$, and $\varepsilon$ which satisfy the false confidence theorem. Specifically, we want to know if $\varepsilon$ can be large enough to have a practically meaningful or harmful effect for statistical inference on $\theta_0$. The relationship between $\varepsilon$ and $p$, for $\alpha = .5$, is plotted in Figure 7.3.1.

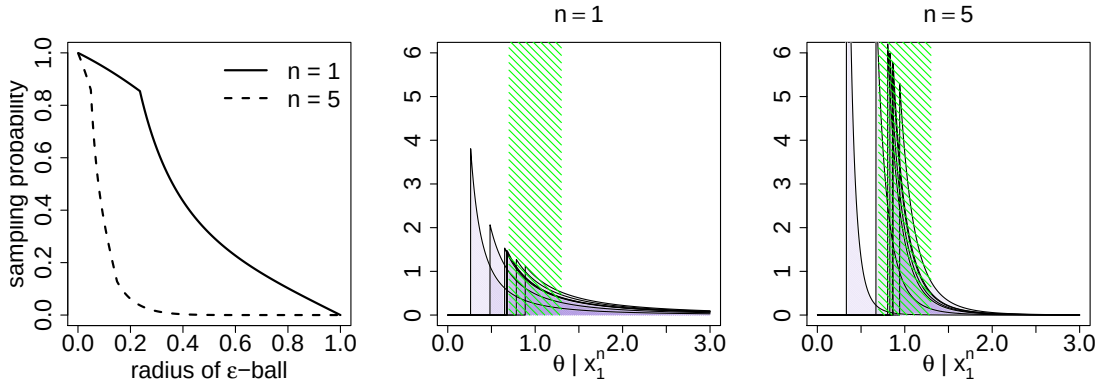



**Figure 7.3.1:** The leftmost panel is a plot of the sampling probability, $p$, as a function of $\varepsilon$, as given by equation (7.3.2), for $\alpha = .5$. The center and rightmost panels are randomly observed realizations of the posterior density of $\theta$, with a .3-ball around $\theta_0$ represented by the shaded green regions. In all panels, the true parameter value is set at $\theta_0 = 1$.

The leftmost panel in Figure 7.3.1 shows, for $\alpha = .5$, the sampling probability (i.e., $p$) that the posterior probability of $A_\varepsilon^c = [\theta_0 - \varepsilon, \theta_0 + \varepsilon]$ is less than $\alpha$, for $\varepsilon$-balls of various radii. For example, with $n = 1$ the posterior probability of $A_\varepsilon^c$ (which contains the true parameter value) will not exceed .5 for $\varepsilon \leq .3$, for more than 80 percent of realized data sets. This has the interpretation that the Bayesian test of "accept $A_\varepsilon^c$" if and only if $P_{\theta|X_1^n}(A_\varepsilon^c) > .5$ would be wrong more than 80 percent of the time.

Displayed on the next two panels of the figure are a few randomly observed realizations of the posterior density of $\theta$, with a .3-ball around $\theta_0$ represented by the shaded green regions. The realizations of the posterior density are typically concentrated around the true value, $\theta_0 = 1$. The next section demonstrates how to extend this example into a situation even more amenable to false confidence.

**Remark 7.3.1.** This uniform example is one of the few simple examples where we can analytically work out the FCT in a straightforward manner. For example, for interval sets, equation (7.3.1) shows the posterior CDF needs an analytic sampling distribution.

### 7.4 Marginal posterior from two uniform distributions

Assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{U}(0, \theta_x)$, and independently $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} \text{U}(0, \theta_y)$. Using the Jeffreys' prior, gives $\theta_x \mid X_1^n \sim \text{Pareto}(n, X_{(n)})$ and $\theta_y \mid Y_1^m \sim \text{Pareto}(m, Y_{(m)})$. Further, define the nonlinear functional $\psi = \theta_x \theta_y$, and derive the posterior distribution of $\Psi$ as follows. By independence,

$$\begin{aligned} P_{\Psi|X_1^n, Y_1^m}(\Psi \leq \psi) &= \int_{Y_{(m)}}^{\infty} P_{\theta_x|X_1^n}\Big(\theta_x \leq \frac{\psi}{\theta_y}\Big) \frac{m Y_{(m)}^m}{\theta_y^{m+1}} \, d\theta_y \\ &= \int_{Y_{(m)}}^{\infty} \Big[1 - \Big(\frac{X_{(n)}\theta_y}{\psi}\Big)^n\Big] \mathbf{1}\Big\{\frac{\psi}{\theta_y} \geq X_{(n)}\Big\} \frac{m Y_{(m)}^m}{\theta_y^{m+1}} \, d\theta_y, \end{aligned}$$

where the last expression results from the form of the Pareto CDF. If $n \neq m$, then this equation simplifies to

$$P_{\Psi|X_1^n, Y_1^m}(\Psi \leq \psi) = 1 + \Big(\frac{m}{n-m}\Big)(X_{(n)} Y_{(m)})^n \psi^{-n} - \Big(\frac{n}{n-m}\Big)(X_{(n)} Y_{(m)})^m \psi^{-m},$$

and if $n = m$, then the distribution function has the form

$$P_{\Psi|X_1^n, Y_1^m}(\Psi \leq \psi) = 1 - \left[1 + n \log\left(\frac{\psi}{X_{(n)} Y_{(n)}}\right)\right] \left(\frac{X_{(n)} Y_{(n)}}{\psi}\right)^n.$$

In both cases, the support of $\Psi$ is $(X_{(n)} Y_{(m)}, \infty)$.

For simplicity, attention will be restricted to the $n = m$ case. This analytic marginal posterior distribution function makes it simple to estimate $p := P_{X_1^n, Y_1^n|\psi_0}\Big(\big\{X_1^n, Y_1^n \colon P_{\psi|X_1^n, Y_1^n}(A_\varepsilon^c) \leq \alpha\big\}\Big)$, for $A_\varepsilon^c = [\psi_0 - \varepsilon, \psi_0 + \varepsilon]$ and various values of $\varepsilon$, by simulating data sets and computing the empirical mean, i.e.,

$$\widehat{p}_k = \frac{\#\big\{X_1^n, Y_1^n \colon P_{\psi|X_1^n, Y_1^n}(A_\varepsilon^c) \leq \alpha\big\}}{k}, \tag{7.4.1}$$

where $k$ is the number of simulated data set pairs $\{X_1^n, Y_1^n\}$. This is done in Figure 7.4.1 for generated data sets. The true values are set at $\theta_x^0 = 10$ and $\theta_y^0 = 1$ which gives $\psi_0 = 10$. Also displayed are a few realizations of the posterior density to illustrate where things go wrong.

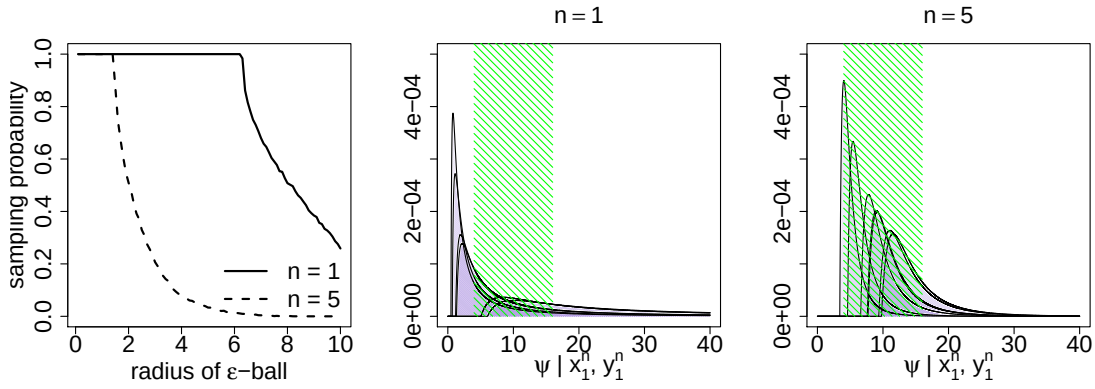



**Figure 7.4.1:** The leftmost panel is a plot of the estimated sampling probability, $\widehat{p}_k$, as a function of $\varepsilon$, as given by equation (7.4.1), for $\alpha = .5$. The center and rightmost panels are randomly observed realizations of the posterior density of $\Psi$, with a 6-ball around $\psi_0$ represented by the shaded green regions. In all panels, the true parameter value is set at $\psi_0 = 10$.

From Figure 7.4.1 it becomes clear how the FCT manifests. For $n = 1$, the $\varepsilon$-ball around $\psi_0 = 10$ with diameter even larger than 12 has posterior probability not exceeding $\alpha = .5$, with sampling probability, $p$, essentially equal to 1. As in the previous section, this has the interpretation that the Bayesian test of "accept $A_\varepsilon^c$" if and only if $P_{\theta|X_1^n}(A_\varepsilon^c) > .5$ would essentially always be wrong. Furthermore, in this case the Bayesian test would fail for an interval (containing the true parameter value) which has length longer than the magnitude of the true parameter value.

Although this is a toy example being used for pedagogical purposes, it is nonetheless alarming. One would hope that the small sample size of $n = 1$, while resulting in less posterior certainty about the location of the true parameter value, would be accompanied by more sampling variability/uncertainty. Rather Figure 7.4.1 implies the interpretation that we are *more* certain about an answer which is in fact false. The center and rightmost panels of Figure 7.4.1 illuminate part of what is happening behind the scene; the posterior densities are typically diffuse around $\psi_0$. The next section presents a more extreme instance of this phenomenon.

### 7.5 Marginal posterior from two Gaussian distributions

Assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{N}(\theta_x, \sigma^2)$, and independently $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{N}(\theta_y, \sigma^2)$. Suppose also that $\sigma$ is known. Using independent Jeffreys' priors, gives $\theta_x \mid X_1^n \sim \text{N}(\bar{X}_n, \sigma^2 n^{-1})$ and $\theta_y \mid Y_1^n \sim \text{N}(\bar{Y}_n, \sigma^2 n^{-1})$. In this context, the nonlinear functional $\psi = \frac{\theta_x}{\theta_y}$ is related to the classical Fieller's theorem in which in-

finite confidence intervals are required to attain frequentist coverage (Fieller, 1954; Gleser and Hwang, 1987; Berger et al., 1999).

The posterior density function for $\psi$ can be derived by transforming the two-dimensional posterior of $(\theta_x, \theta_y)$ into the space of $(\psi, \gamma) = (\frac{\theta_x}{\theta_y}, \theta_y) =: g(\theta_x, \theta_y)$ and then computing the marginal distribution of $\psi$. Observe that $g^{-1}(\psi, \gamma) = (\psi\gamma, \gamma)$ which gives the Jacobian for the transformation,

$$J(\psi, \gamma) = \det \begin{pmatrix} \gamma & \psi \\ 0 & 1 \end{pmatrix} = \gamma.$$

Then the joint posterior density has the form

$$\pi_{\psi,\gamma|X_1^n,Y_1^n}(\psi, \gamma) = \pi_{\theta_x|X_1^n}(\psi\gamma) \cdot \pi_{\theta_y|Y_1^n}(\gamma) \cdot |\gamma| \cdot \mathbf{1}\{\gamma \neq 0\}.$$

Recalling the forms of the posterior densities for $\theta_x$ and $\theta_y$, and integrating over $\gamma$ gives

$$\begin{aligned} \pi_{\psi|X_1^n,Y_1^n}(\psi) &= \int \pi_{\psi,\gamma|X_1^n,Y_1^n}(\psi, \gamma)\, d\gamma \\ &= \left(\frac{n}{2\pi\sigma^2(1+\psi^2)}\right)^{\frac{1}{2}} \exp\left\{\frac{n}{2\sigma^2}\left[\frac{(\psi\bar{X}_n + \bar{Y}_n)^2}{1+\psi^2} - \bar{X}_n^2 - \bar{Y}_n^2\right]\right\} \cdot E_{\gamma|\psi}(|\gamma|), \end{aligned} \tag{7.5.1}$$

where the expectation is taken over $\gamma \mid \psi \sim \mathrm{N}\left(\frac{\psi\bar{X}_n + \bar{Y}_n}{1+\psi^2}, \frac{\sigma^2}{n(1+\psi^2)}\right)$.

This marginal posterior is easily estimable, and $p := P_{X_1^n,Y_1^n|\psi_0}\left(\{X_1^n, Y_1^n : P_{\psi|X_1^n,Y_1^n}(A_\varepsilon^c) \leq \alpha\}\right)$, for $A_\varepsilon^c = [\psi_0 - \varepsilon, \psi_0 + \varepsilon]$ and various values of $\varepsilon$, can be estimated with an approximating Riemann sum using equation (7.5.1). The estimated $p$ as a function of $\varepsilon$ is displayed in Figures 7.5.1 and 7.5.2 for $\alpha = .5$ and $\alpha = .05$, respectfully, and for various noise levels, $\sigma$. The true mean values are set at $\theta_x^0 = .1$ and $\theta_y^0 = .01$ which gives $\psi_0 = 10$. Displayed in Figure 7.5.3 are a few random realizations of the posterior densities from (7.5.1), for various sample sizes, $n$, with $\sigma = 1$, to illustrate part of where things go wrong.

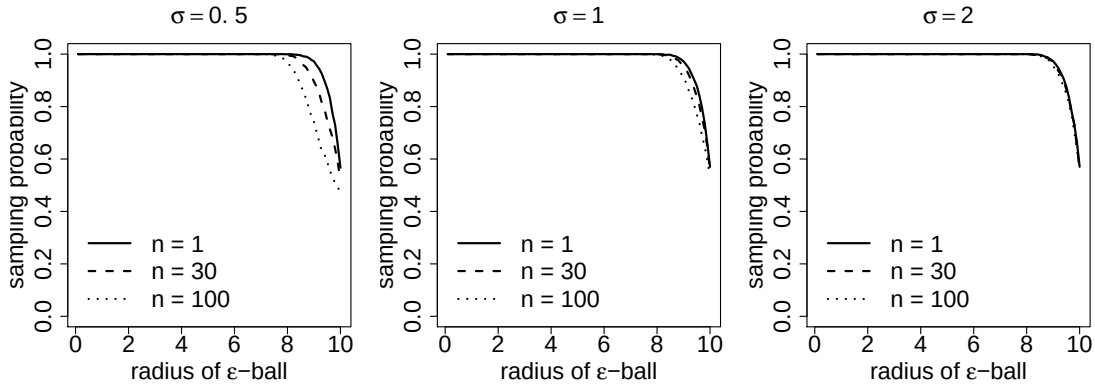



**Figure 7.5.1:** Each panel is a plot of the estimated sampling probability of $p$, as a function of $\varepsilon$, using the posterior density equation (7.5.1), and setting $\alpha = .5$. The true parameter value is $\psi_0 = 10$.

Remarkably, for almost all values of $n$ and $\sigma$ considered in Figure 7.5.1 the Bayesian test of "accept $A_\varepsilon^c$" if and only if $P_{\theta|X_1^n}(A_\varepsilon^c) > .5$ would fail for $\varepsilon$ as large as 8. Even considering the extreme choice of $\alpha = .05$ as in Figure 7.5.2, the sampling probability, $p$, exceeds 80 percent chance (in the case of $\sigma = 1$) that $P_{\theta|X_1^n}(A_\varepsilon^c) \leq .05$ for $\varepsilon$ as large as 4, with $n = 100$.

A further illustration of what is happening is once again provided with random realizations of the marginal posterior densities presented in Figure 7.5.3. For this problem they heavily concentrate away from the true value $\psi_0 = 10$. Consequentially, any inference on the true value of $\psi$ is sure to be misleading, and hence this situation is an extreme example of the manifestation of false confidence in a well-studied classical problem. Similar results hold for the manifestation of false confidence in other non-linear marginalization examples, e.g., the coefficient of variation which is discussed in the Appendix.

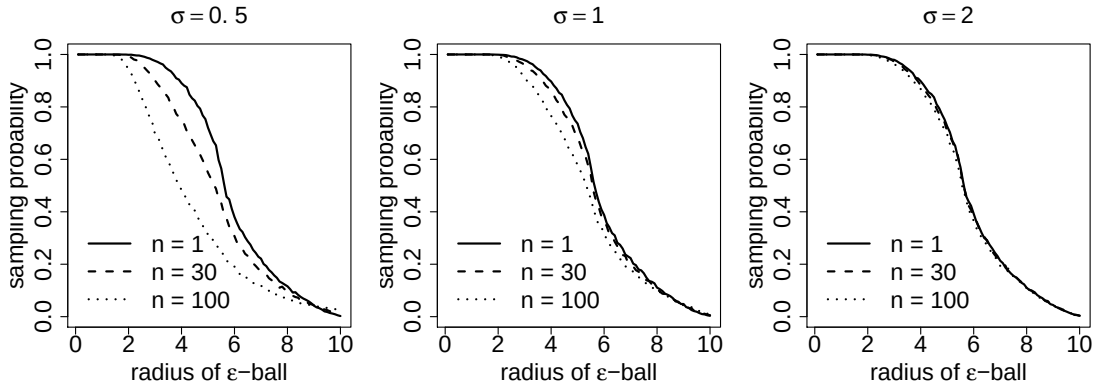



**Figure 7.5.2:** Each panel is a plot of the estimated sampling probability of $p$, as a function of $\varepsilon$, using the posterior density equation (7.5.1), and setting $\alpha = .05$. The true parameter value is $\psi_0 = 10$.

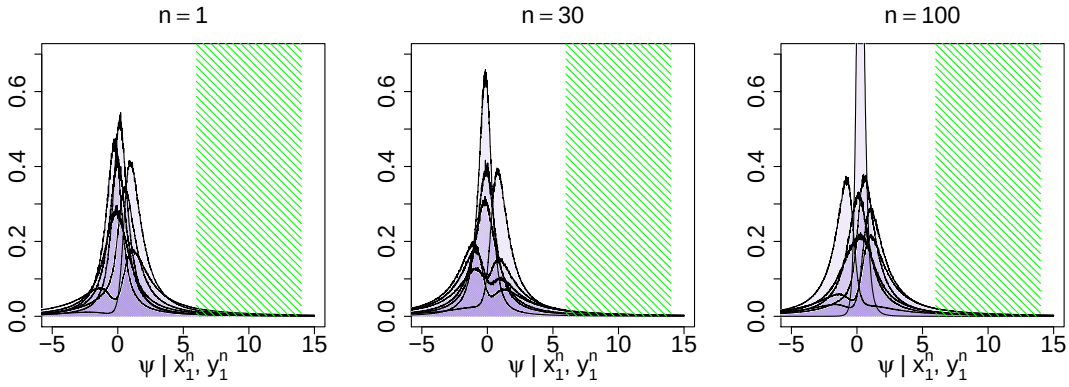



**Figure 7.5.3:** Each panel exhibits randomly observed realizations of the posterior density of $\psi$, equation (7.5.1), with a 4-ball around $\psi_0 = 10$ represented by the shaded green regions.

### 7.6 Concluding remarks and future work

There is currently little theoretical understanding of the phenomenon of false confidence or of when it plays a significant role in statistical analysis. We demonstrate ramifications of false confidence in standard, single parameter models as well as models involving the marginalization of multiple parameters. Our examples illustrate that models involving the marginalization to non-linear, not one-to-one functions of multiple parameters seem to play a key role in more extreme manifestations of false confidence. In future work we seek to gain an understanding of why the FCT is problematic in these situations.

### 7.7 Acknowledgments

The authors are grateful to Ryan Martin, Jan Hannig, and Samopriya Basu for many helpful comments, engaging conversations, and encouragement.

CHAPTER 8

# Data science vs. statistics: two cultures?

Data science is the business of learning from data, which is traditionally the business of statistics. The term data science, however, is often viewed as a broader, task-driven and computationally-oriented version of statistics. Examining and expanding upon the views of a number of statisticians, this chapter encourages a big-tent view of data analysis. Data science, both the term and idea, has its origins in statistics, representing a reaction to a narrower view of data analysis. We examine how different and evolving approaches to data analysis are related to broader trends in data science (e.g. exploratory analysis, machine learning, reproducibility, computation, etc). Finally, we discuss how these trends relate to academic statistics, including future directions for communication, education and research.

## 8.1 Introduction

A simple definition of a data scientist is someone who uses data to solve problems. In the past few years this term has caused a lot of buzz[1] in industry, questions in science[2] and consternation in statistics[3]. One might argue that *data science* is simply a rebranding of *statistics* (e.g. "data science is statistics on a Mac"[4]), but this comment misses the point.

While data analysis has been around for a long time, its economics (costs of and value derived) have changed primarily due to technology driving: the availability of data, computational capabilities and ease of communication. The most obvious advances are in computer hardware (e.g. faster CPUs, smaller microchips, GPUs, distributed computing). Similarly, algorithmic advances play a big role in making computation faster and cheaper (e.g. optimization and computational linear algebra). There are many new/improving technologies which allow us to gather data in new, faster and cheaper ways, including:

[1] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[2] https://www.wired.com/2008/06/pb-theory/

[3] https://simplystatistics.org/2015/10/29/the-statistics-identity-crisis-am-i-really-a-data-scientist/

[4] https://twitter.com/cdixon/status/428914681911070720

drones, medical imaging, sensors (e.g. Lidar), better robotics, Amazon's Mechanical Turk, wearables, etc. Improved software (e.g. see Section 8.4.2.1) makes it faster, cheaper and easier to: communicate the results of data analysis, distribute software, and publish research/educational resources.

These changes mean that more people get more value out of data analysis. One would be hard pressed to find an industry or academic discipline where data analysis is not having an impact. These developments also mean that data analysis has become increasingly multidisciplinary and collaborative. The field of data analysis has broadened in that more people want to analyze data and data analysis draws on more disciplines. Areas of statistics previously considered specialized (e.g. statistical software, exploratory analysis, data visualization, high dimensional analysis, complex data objects and the use of optimization methods) have become dramatically more valuable. All of these changes have an impact on the discipline of statistics.

Various viewpoints have been expressed on the current state/future of statistics and the fashionable topic of *data science* (Tukey, 1962; Wu, 1998; Cleveland, 2001; Breiman et al., 2001; Hand et al., 2006; Yu, 2014; Wasserman, 2014; Efron and Hastie, 2016; Donoho, 2017; Hooker and Hooker, 2017; Bühlmann and Stuart, 2016; Blei and Smyth, 2017; Barocas et al., 2017; Bühlmann and van de Geer, 2018; Reid, 2018). Views have also been been expressed in blogs: Simply Statistics[5], Statistical Modeling, Causal Inference, and Social Science[6], and Normal Deviate[7].

This chapter offers additional opinions and perspectives and we strive for a unique viewpoint through combining both new and old ways of thinking. The first author is a computationally oriented PhD student in a statistics, who has developed and taught an entirely new undergraduate course on data science[8] for his department. The second author has 35 years of academic experience in statistics including: research, collaborative applications and teaching at all levels. He has offered previous opinions on *big data*, in particular on the topic of *robustness against heterogeneity* in (Marron, 2017).

In Section 8.2 we discuss the definition of *data science* and how it relates to the discipline of *statistics*. In Section 8.3 we discuss a few *modes of variation* of data analysis which give insights into aspects of modern data analysis (e.g. why the rise of computation is tied to the rise of exploratory and predic-

[5] https://simplystatistics.org/

[6] http://andrewgelman.com/

[7] https://normaldeviate.wordpress.com/

[8] https://idc9.github.io/stor390/

tive analysis). Finally, in Section 8.4 we discuss future directions of statistics research, communication and education including: complex data, robustness, machine learning and data processing, and literate programming.

### 8.2 What is Data Science

this chapter

To a general audience, data science is often defined as the intersection[9] of three areas: math/statistics, computation and a particular domain (e.g. biology) (Conway, 2010; Yu, 2014; Blei and Smyth, 2017). Implicit in this definition is the focus on solving specific problems (in contrast with the type of deep understanding that is typical in academic statistics[10]). The focus on problem solving is important because it explains differing judgements to be found on how to value contributions to the field. As stated in (Cleveland, 2001)

> [results in] data science should be judged by the extent to which they enable the analyst to learn from data... Tools that are used by the data analyst are of direct benefit. Theories that serve as a basis for developing tools are of indirect benefit.

For the purpose of this article we define data science as the union of six areas of *greater data science* which are borrowed from David Donoho's article titled "50 Years of Data Science" (Donoho, 2017). We will refer to these as GDS 1-6 and more details can be found in Section 8 from Donoho's article[11].

1. Data gathering, preparation, and exploration
2. Data representation and transformation[12]
3. Computing with data
4. Data modeling
5. Data visualization and presentation

[9] This idea is usually communicated through a venn diagram e.g. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

[10] https://simplystatistics.org/2014/07/25/academic-statisticians-there-is-no-shame-in-developing-statistical-soluti

[11] http://www.tandfonline.com/doi/pdf/10.1080/10618600.2017.1384734

[12] This includes both databases and mathematical representations of data.

6. Science about data science[13]

The purpose of defining data science in this way is to (A) better capture where people who work with data spend their time/effort and (B) put more focus on the value of each tool for providing insights. This definition is given in contrast to the current, perceived state of statistics. A number of other statisticians have proposed similar definitions using different terminologies and ontologies (Tukey, 1962; Chambers, 1993; Cleveland, 2001; Yu, 2014). People might reasonably tweak the above definition.

The term *data science*, both the literal string and the broader idea that it conveys, originates from statisticians (see Section 8.2.1). In a 1993 essay titled "Greater or Lesser Statistics: a Choice for Future Research" statistician John Chambers wrote (Chambers, 1993)

> Greater statistics can be defined simply, if loosely, as everything related to learning from data, from the first planning or collection to the last presentation or report. Lesser statistics is the body of specifically statistical methodology that has evolved within the profession - roughly, statistics as defined by texts, journals, and doctoral dissertations. Greater statistics tend to be inclusive, eclectic with respect to methodology, closely associated with other disciplines, and practiced by many outside of academia and often outside professional statistics. Lesser statistics tends to be exclusive, oriented to mathematical techniques, less frequently collaborative with other disciplines, and primarily practiced by members of university departments of statistics.

We take the position that data science is a reaction to the narrow understanding of *lesser statistics*; simply put, data science has come to mean a broader view of statistics.

It's worth noting that our discussion is about data science from the perspective of statistics. Since data science is a multidisciplinary field, other disciplines, such as computer science, might see data science in a different way. For example, computer science might focus more on: machine learning, large scale computation and data storage/retrieval.

[13] For example, reproducible research would fall under this category and point 5.

### 8.2.1 What's in a Name?

The term data science has caused excitement, confusion and controversy. Some of the confusion is from the lack a of a consistent definition[14]. There is an ecosystem of related terms (e.g. *analytics, business intelligence, big data, data mining*). Many companies/people/organizations have their own internal definition for each of these terms e.g. one company's data scientist is another company's business analyst. The lack of a consistent term makes discussion challenging.

These terms are contentious partially because of the buzz associated with them and because of arguments about academic discipline subsetting e.g. some argue data science is a subset of statistics[15] while others argue statistics is a subset of data science[16]. Academic funding (and egos) play a non-trivial role in the controversy.

Even the origin of the term is up for debate. We point to a few sources for developing[17] both the literal string and the broader idea, "data science"

- (Tukey, 1962): the idea, not the literal string
- (Naur, 1974): the literal string, not the idea[18]
- (Chambers, 1993): the idea, not the literal string
- (Wu, 1998): both the idea and literal string
- (Cleveland, 2001): both the idea and literal string

Others are credited with bringing the term/idea to industry e.g. (Patil, 2011). For more on the development of the term see (Donoho, 2017) and the two articles linked to below[19]. Both the idea (as defined in Section 8.2) and the literal string (A) have been around for a while and (B) have origins in statistics.

The term data science has broken free from academic statistics into industry and other academic fields. In some cases, data science marginalizes the discipline of statistics which is a detriment to both

[14] We take the position that data science is the practice of broader statistics.

[15] "a data scientist is a statistician who lives in San Francisco" (Bhardwaj, 2017)

[16] http://andrewgelman.com/2013/11/14/statistics-least-important-part-data-science/

[17] We don't claim this list is exhaustive.

[18] Peter Naur uses the term "data science" but in a narrower sense, focusing more on computation.

[19] http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/ and https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

statistics and anyone who analyzes data. Acknowledging the history of data science and statistics we hope will garner more respect for data science within statistics and for statistics from the broader community of data practitioners.

In the next few section we discuss the origins of data science, critiques of statistics and the broader notion of reproducibility.

### 8.2.2 Critiques of Statistics

To first order, we summarize the critiques of statistics as: too much theory, not enough computation. We believe theory is important, however, too much theory at the expense of other things is a detriment to the field. See Section 8.4 for a more optimistic discussion of theory.

We thank Hongtu Zhu for pointing out two quotes from the Priceonomics website (Bhardwaj, 2017)[20], which provides some interesting discussion about data science and statistics.

> Statistics was primarily developed to help people deal with pre-computer data problems like testing the impact of fertilizer in agriculture, or figuring out the accuracy of an estimate from a small sample. Data science emphasizes the data problems of the 21st Century, like accessing information from large databases, writing code to manipulate data, and visualizing data.

This quote is echoed by statisticians such as Hadley Wickham who lament the lack of priority academic statistics has given to such areas[21]. This point is well taken e.g. what proportion of statistics undergraduates are competent in R or Python? However, statistics is being sold short on its contributions to computation (Donoho, 2017). For example, the R-Project (Members, 2017) is a context where many members of the statistical community are directly engaged in writing and sharing code. Furthermore, there is a rather large area called *statistical computing*, with quite a long history, see e.g. the American Statistical Association's Section[22], which has been in continuous operation since 1972.

Another example is visualizing data. We acknowledge that a large fraction of the statistical community has created a culture of not devoting enough energy in the direction of looking at data. However,

[20] https://priceonomics.com/whats-the-difference-between-data-science-and/

[21] E.g. see https://priceonomics.com/hadley-wickham-the-man-who-revolutionized-r/ and the quote about "The fact that data science exists as a field is a colossal failure of statistics."

[22] http://stat-computing.org/computing/

again there is a relatively small but very active community devoted to visualization, including the American Statistical Association's Section on Graphics[23], formed in 1985.

Another popular, but incorrect belief is that statistics is not concerned with *big data* (or phrased sans buzz words: statisticians do not care about computing things efficiently). As (Donoho, 2017) points out, statisticians have in fact always been interested in large data computation. For example, the word *statistics* came about from work on census data which have been around for centuries and are large even by today's standards. The principle of *sufficiency* is of course a mechanism to deal with large data sets efficiently. The point here is that these pursuits are a part of statistics, but are perhaps considered specialized as opposed to mainstream (e.g. in terms of publications in flagship journals, undergrad/graduate education, etc).

A second quote from Priceonomics (also echoed by (Chambers, 1993; Donoho, 2017))

> Statistics, on the other hand, has not changed significantly in response to new technology. The field continues to emphasize theory, and introductory statistics courses focus more on hypothesis testing than statistical computing... For the most part, statisticians chose not take on the data problems of the computer age.

This point is on target in a number of ways, but we take a different viewpoint on a few issues.

The first is on the value of statistical theory. Far from viewing it as something old fashioned and hence useless, we argue that the need for theoretical thinking is greater than ever in the data science age (see Section 8.4). There are few calls yet for *data science theory*, although the US National Science Foundation's TRIPODS program[24] is an important exception. However we predict that as the field evolves there will be growing realization of the importance of that line of thought, as more and more algorithms become available with very little meaningful basis for making the critical choice among them.

The second is the statement that "statistics courses focus more on hypothesis testing". This makes the statistical outsider's mistake of thinking that statistics is a set of recipes for doing data analysis. It misses the deeper truth understood by people who practice data analysis: when properly taught, statistics courses teach an important way of thinking called the *scientific method*. The main idea is that to be

[23] http://stat-graphics.org/graphics/

[24] https://www.nsf.gov/funding/pgmsumm.jsp/pimsid=505347

really sure of making actual discoveries (as opposed to finding spurious and ungeneralizable sampling artifacts) scientists should first formulate a hypothesis, then collect data and finally analyze.

One can be forgiven, however, for mistaking statistics as a set of recipes. Too many people interact with statistics exclusively via a standard Statistics 101 type class which may in fact treat statistics as a handful of formulas to memorize and steps to follow. While we believe the material taught in these courses is vital to doing science, it is perhaps time to rethink such introductory classes and teach data before (or concurrently with) teaching statistics. See Section 8.4.3 and the references therein.

### 8.2.3 Reproducibility and Communication

The idea of *reproducibility* has become a hot topic recently in science (Peng, 2011; Stodden, 2012; Kiar et al., 2017). A narrow sense of reproducibility is *the ability to recompute results* i.e. that someone else can easily obtain the data and run the code used in the original experiment (Leek and Peng, 2015; Patil et al., 2016). Reproducibility, in the broader sense, is about three things: scientific validity, communication and methodological development.

A number of factors go into assessing the scientific validity of a study. The gold standard of scientific validity is *independent replication*. Replicability is about the ability for someone else to rerun the same experiment and obtain the same results[25]. Reproducibility is about checking the details of the scientific argument made in a paper (much in the way a mathematician would check the details of a proof of a theorem). The requirement that the code which produced the results can be rerun at a later time and will create the exact same figures/numbers is an important condition for assessing the correctness of the results. It is also a software engineering issue; the scientist has to write code which is understandable, well documented, publicly available, and persists across time and computers[26].

The analysis code is a key part of communicating the analysis itself, citing Jon Claerbout (Buckheit and Donoho, 1995)

> [a]n article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete soft-

[25] The literature is not consistent about the definitions of reproducibility and replicability. In this chapter we use the definitions given here.

[26] Writing code that continues to work overtime is non-trivial; it involves maintaining the same computing environment and managing dependencies correctly e.g. the software packages the code uses change over time, version 1.1.1 might work the same as version 2.1.1.

ware development environment and the complete set of instructions which generated the figures.

The text summary of an analysis in the paper may not include all the details of a complex analysis (e.g. how were hyper-parameters tuned, how were the data preprocessed, etc). Moreover the description of the analysis may be incorrect: numbers can be misreported by accident or intentionally, the code may not work the way the analyst thought it did[27], there may be bugs in the code, etc.

Computational methodologies are often reused and built upon. If the code for a methodology is not available then the next person who wants to use/iterate on that methodology has to re-write the code[28] which is both an inefficient use of time/resources and can lead to errors.

Three of the main barriers to reproducibility are: culture, computational tools and education (Peng, 2011; Stodden, 2012; Leek and Peng, 2015). Even if the cultural and educational problems are solved, there are still technical challenges which discourage reproducibility in practice. These issues are primarily about software engineering; while reproducibility is technically achievable, it is often too burdensome in practice[29]. For example, "in a recent survey of the machine learning community–the single biggest barrier to sharing code and data was the time it takes to clean up and document the work to prepare it for release and reuse" (Stodden, 2012). Other issues include: data sharing, data privacy, software sharing, software verification (e.g. unit testing), writing readable code, version control, and legality.

Section 8.4.2 discusses how the rise of *literate programming* in data science has improved our ability to do and communicate reproducible science. It also discusses how the sharing of code and educational resources is a boon to the field.

### 8.3 Modes of Variation

In this section we discuss different approaches to data analysis and how they relate to broader trends in data science. We call these *modes of variation* in the sense that they explain variation in the practice of data science. Each of the modes discussed below (e.g. predictive vs. inferential analysis) represents a

[27] Understanding the nitty-gritty details of how statistical software works is not trivial: how does the optimization routine determine convergence? Are the data mean centered by default? There is a lack in uniformity in how statistical software is written; we believe this is exacerbated by the lack of of statisticians writing statistical software.

[28] Even if the code for a study is available, someone may still want to rewrite the code say in another language. In this case have the original code available to base the new code on is helpful.

[29] Publishing messy code is still beneficial and certainly better than not publishing code (Barnes, 2010)

spectrum between two methodologies, one more associated with data science and the other with classical statistics. These modes help explain why different communities seem to talk past each other and why some techniques (e.g. computation) have become more popular in recent times.

### 8.3.1 Prediction vs. Inference (Do vs. Understand)

A computer scientist might pejoratively describe a linear or logistic regression as shallow and quaint[30]. A statistician might express bewilderment at the buzz around deep learning and question why more principled, interpretable statistical models won't do the trick. The point here is that these two imaginary, curmudgeonly academics are thinking about problems with different goals. The computer scientist is trying to build a system to accomplish some task; the statistician is typically trying to learn something about how the world works.

Prediction vs. inference is a spectrum. Many complicated problems have well defined subproblems which are closer to one end or the other end. The distinction we are trying to make here is maybe better described as engineering vs. science (or at least broad generalizations thereof). Engineering is the business of creating a thing that *does* something[31]. Science is the business of *understanding* how something works. Of course engineers use science and scientists use engineering. But if we focus on the end goal of a particular problem we can probably, reasonably classify that problem as either science or engineering[32].

(Breiman et al., 2001) discusses many of the differences between predictive and inferential modeling. Predictive modeling often uses more sophisticated, computationally intensive models. This often comes with a loss in interpretability and general understanding about how the model works and what the data look like (Freitas, 2014). Predictive modeling also places less emphasis on theory/assumptions because there are fairly good, external metrics to tell the analyst how well they are doing (e.g. test set error).

Predictive modeling is one of the main drivers of *artificial intelligence* (AI). The fact that data can be used to help computers automate things is perhaps one of the most impactful innovations of recent decades. Early attempts at AI type applications involved primarily *rules based systems* which did not use

[30] The use of shallow means we can view a generalized linear model as a neural network with 0 layers. The more layers a network has, the more complex of a pattern it can find (Goodfellow et al., 2016).

[31] I.e. the output of a predictive model may be interesting insofar as it helps us do something.

[32] In other words, in many cases understanding is primarily a means to and ends for predictive problems and visa versa.

a lot of data (Russell and Norvig, 2009); modern AI systems are typically based on *deep learning* and are extremely data hungry (Goodfellow et al., 2016). Section 8.4.1.4 discusses some of the ways statisticians can make important contributions to AI/predictive modeling. Section 8.4.1.5 discuss some of the ways statisticians/scientists may benefit from and reasons to be aware of machine learning.

### 8.3.2 Empirically vs. Theoretically driven

> Data science is exploratory data analysis gone mad. – Neil Lawrence[33]

Most quantitative fields of study do both theoretical and empirical work[34] e.g. theoretical vs. experimental physics. Within statistics, we might contrast *exploratory data analysis* (EDA) vs. *confirmatory analysis* i.e. searching for hypotheses vs. attempting to confirm a hypothesis (in this section *theory* refers to the scientific theory being tested).

It used to be that statistics and science were primarily theory driven. A scientist has a model of the world; they design and conduct an experiment to assess this model; then use hypothesis testing to confirm the results of the experiment. With changes in data availability and computing, the value of exploratory analysis, data mining, and using data to generate hypotheses has increased dramatically (Fayyad et al., 1996; Blei and Smyth, 2017). EDA often prioritizes the ability to rapidly experiment which means computation can dominate the analysis.

EDA can become problematic when the analyst puts too much faith in its results i.e. when the analyst mistakes hypothesis generation for hypothesis confirmation. Every statistician can list problems with simply *correlation mining* a data set (e.g. false discovery, sampling bias, etc). These problems don't mean the results are wrong, but rather they mean exploratory analysis provides much weaker evidence for a hypothesis than confirmatory analysis. The amount this matters is context dependent.

One problematic idea is that EDA can solve every problem. For example, in the controversially titled article, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" it was argued that EDA will replace the scientific method (Anderson, 2008). We disagree. This article is an extreme example of the broader attitude that correlation, and fancy models applied to large data sets, can replace causal

[33] From Talking Machines season 3, episode 5 https://www.thetalkingmachines.com/.

[34] This statement probably applies to non-quantitative fields. For example, some academics in comparative literature are more "empirical" in the sense they examine a particular body of work, draw conclusions and possibly generalize/relate their conclusions to other bodies of work. Other people in comparative literature apply "theoretical methods."

inference and the careful, time intensive scientific method. The point that EDA can contribute to scientific applications is well taken, however, and this in fact is becoming more common (Blei and Smyth, 2017). These applications likely raise many interesting and impactful research questions arising from the increased value in EDA (in all of GDS 1-6).

### 8.3.3 Problem First vs. Hammer Looking for a Nail

Some researchers take a *hammer looking for a nail* approach; the researcher has developed/studied a statistical procedure and then looks for problems where it might be applicable. Other researchers aim to solve some particular problem from a domain. Note that the former approach is strongly correlated with, but not equivalent to theoretical research (same for the latter and applied research).

Both research approaches are valid and productive, however the balance in academic statistics may have shifted too far to the former (hammer) approach. For example, Rafael Irizarry has some interesting further commentary[35] which is echoed by a number of others (Tukey, 1962; Breiman et al., 2001; Wasserman, 2014; Donoho, 2017). We include this section because data science is focused on problem solving and it is this problem solving which makes data analysis useful to other disciplines.

### 8.3.4 The 80/20 Rule

The 80/20 rule of data analysis is:[36]

> One of the under appreciated rules in statistical methods development is what I call the 80/20 rule (maybe could even by the 90/10 rule). The basic idea is that the first reasonable thing you can do to a set of data often is 80% of the way to the optimal solution. Everything after that is working on getting the last 20%.

Applying basic models to a data set often provides the most value (and/or solves the problem of interest much of the time). The 80/20 rule explains part of why a number of techniques have become more valuable than in the past and why the six areas of GDS emphasize previously undervalued areas. These include: data visualization, exploratory data analysis, data mining, programming, data storage/processing, computation with large datasets and communication.

[35] https://simplystatistics.org/2014/07/25/academic-statisticians-there-is-no-shame-in-developing-statistical-soluti

[36] https://simplystatistics.org/2014/03/20/the-8020-rule-of-statistical-methods-development/

## 8.4 Going Forward

In this section we discuss a number of areas we believe are particularly promising for statistics research, communication and education. A pervasive theme in this section is the role of computation, defined broadly as in (Nolan and Temple Lang, 2010).

### 8.4.1 Research

As data analysis becomes more valuable, existing statistical theory and methodology also become more valuable. Criticisms of statistical theory (e.g. Section 8.2.2 and (Donoho, 2017)) are largely about ignoring other, less mathematically glamorous areas of statistics.

In this section we highlight a number of (potentially) promising areas of statistics research. There are many ways in which one can do *good* statistical research, including all of those discussed in (Tao, 2007). The list below is biased in favor of our tastes and areas which involve both theory and aspects of greater data science (from Section 8.2). Rather than diminishing the role of mathematical statistics, we emphasize that many of these areas will require novel contributions from mathematical statistics which itself should be broadened.

#### 8.4.1.1 Complex Data and Representation

The recently fashionable area of Big Data has received a large amount of well deserved attention and has led to many serious analytic challenges. Currently less well understood is that a possibly bigger challenge is non-standard or *complex data*. In particular, many modern data analytic scenarios involve non-standard data such as: networks, text, sound, images, movies, shapes, very high dimensional data, data living on a manifold, etc. In response to this challenge, (Wang and Marron, 2007; Marron and Alonso, 2014) have proposed *Object Oriented Data Analysis* (OODA).

A fundamental concept of OODA is that in complex scientific contexts, it is often not even clear what should be the atoms of the statistical part of the analysis, i.e. the experimental units. OODA provides terminology for pivotal team discussions on this topic between domain scientists and statisticians which lead to an effective final analysis, via resolving *what should be the data objects?* An interesting example of this in an image analysis context can be found in (Lu et al., 2014).

The OODA discussion goes beyond choice of experimental units also to include the key technical issue of *data representation*. This includes both fairly standard statistical techniques such as data transformation, but also mathematically deeper issues such as the appropriate data space, as discussed for example in (Pizer and Marron, 2017). Much of the success of deep learning may come from its ability to automatically discover useful data representation (Bengio et al., 2013); connections between OODA and *representation learning* are unexplored and potentially fruitful directions for future research.

Complex data and OODA present many research opportunities such as: invention of powerful new tools for data practitioners, computational challenges, methodological developments, and developments in statistical theory. OODA often involves bringing in a number of mathematical disciplines such as differential geometry, topology, optimization, probability, etc, providing a sense in which data science should become a truly interdisciplinary endeavor. It is also an opportunity to greatly extend usage of the term *mathematical statistics* to include many more mathematical areas beyond just the conventional probability theory.

#### 8.4.1.2 Robustness to Unknown Heterogeneity

An even less widely acknowledged and studied challenge in data science is *data heterogeneity*. This topic is very current to many modern sciences, where there has been a growing realization that complex data collection by a single lab tends to result in sample sizes that are inadequate to address many modern scientific issues of interest.

The poster child for this problem may be cancer research, where the very diversity of the disease requires very large sample sizes in a context where the needed measurements are very labor, time and cost intensive. Such challenges have led to the formation of large multi-lab research consortiums. In the cancer world, a well known effort of this type is The Cancer Genome Atlas (TCGA) (Network et al., 2012). While great care is taken in such efforts to try to standardize laboratory protocols and many other aspects of the data collection, it is well known among all data collectors that there are always impossible to control biases that creep into the data set when data from different labs are combined.

Dealing with this issue is the main challenge of data heterogeneity. It is clear that the standard statistical Gaussian model for noisy data is no longer appropriate. When data from different sources are combined, a much more appropriate statistical model is a Gaussian mixture, but this presents a major challenge to classical statistical approaches such as the likelihood principle.

Much more research is strongly needed on all aspects of data analysis for heterogeneous data, including methods, computation and theory. It is also clear that real progress is going to be made by approaches which truly integrate all three of those. More discussion of important early work in that direction can be found in (Bühlmann and Meinshausen, 2016; Bühlmann and Stuart, 2016; Marron, 2017).

#### 8.4.1.3 Scalable and Robust Models

Robustness issues are important for modern, large data applications (see (Huber, 2011; Hampel et al., 2011; Staudte and Sheather, 2011; Maronna et al., 2006) for an overview of this area). For the gains to be realized from robust statistical procedures for large data sets, these procedures need to be computationally efficient (or *scalable*). We want models which are both scalable and robust. We point to (Bottou et al., 2016) for an overview of important computational methods in machine learning.

Robustness presents a challenge for developing scalable models for large data because often robust methods are harder to compute[37]. Models which are both scalable and robust present an opportunity for research developments drawing on all of statistics, optimization and computer science. The work of (Aravkin and Davis, 2016) is a recent example in this vein. We point this area out because it requires joint reasoning based on both computation and statistical theory.

#### 8.4.1.4 Automation and Interpretability

One of the biggest ways data is impacting society is by powering automation through machine learning. While data driven automation has a lot of potential to do good, recent years have brought new attention to its negative consequences (Zarsky, 2016; O'Neil, 2017; Doshi-Velez and Kim, 2017; Crawford, 2017). For example, (Bolukbasi et al., 2016) demonstrate that natural language processing algorithms trained on google news text data learned offensive gender stereotypes. In a book titled "Weapons of Math Destruction" data scientist Cathy O'Neil provides many examples of how automation can perpetuate and even reinforce existing societal inequalities (O'Neil, 2017).

Statistics is the discipline historically most concerned with the myriad of ways in which data can be misleading. The rise of automation presents new opportunities to the discipline. First, the decades of statistics' hard won knowledge about dealing with data is salient to many applications of data driven

[37] E.g. L2 regularized (ridge) linear regression has a closed form while L1 regularization (LASSO) does not.

automation. Second, automation presents new technical challenges to statistics since machine learning often involves applying sophisticated modeling techniques to large, complex datasets. One of the primary technical challenges is interpretability.

*Interpretability*, as discussed in (Lipton, 2016; Doshi-Velez and Kim, 2017), is desirable for a number of reasons including: trust, causality, transferability, informativeness, fair and ethical decision-making, legality, and safety. Additionally, interpretability can mean different things e.g. a visualization, a verbal explanation, understanding the model, etc. The canonical examples of interpretable models are generalized linear models and decision trees. The canonical example of an *uninterpretable* model is a deep neural network. One interesting approach to interpretability is discussed in (Ribeiro et al., 2016), called *local interpretable, model-agnostic explanations* (LIME), which proposes the use of simple models to help explain predictions made by more sophisticated models.

The pursuit of understanding machine learning models to make them more interpretable is a wide open area for statisticians. These questions may involve narrowing the gap between predictive and inferential modeling discussed in Section 8.3.1.

#### 8.4.1.5 Machine Learning and Data Processing

Complex data preprocessing pipelines are familiar to statisticians working in many areas including genomics and neuroscience (Gentleman et al., 2006; Kiar et al., 2017). The raw data go through a number of processing steps before reaching the analyst as a .csv file. These steps may include: simple transformations, complex algorithms, output of statistical models, etc. Knowledge of the processing pipelines can be important for the analyst because the pipelines can have errors and introduce systemic biases in the data.

Machine learning, particularly deep learning, may cause complex data processing pipelines to be more frequently used in modern data analysis. A lot of work in machine learning is about measurement: image recognition, image captioning, speech recognition, machine translation, syntactic parsing, sentiment analysis, etc (Goodfellow et al., 2016). These examples some what blur the line between data processing and data gathering. We suspect that it will be more common for data analysts to have one or more variables in their datasets which are the output of a deep learning model. While machine learning can provide the analyst with rich, new variables, it should also give the statistician some pause.

There are at least two major issues with using a deep learning model to gather data: it puts black box models into the data processing pipeline and introduces an external data set (i.e. the one used to train the deep learning model). Neither of these are new issues, but they will become more prominent. The black box model is problematic because it means the analyst cannot fully understand where the raw data came from. External data is problematic because it can infect new datasets it comes into contact with via the deep learning model on which it was trained[38]. Both of these issues are likely to create systemic biases which are challenging to detect and appropriately handle.

We suspect that machine learning will create a lot of value in data processing in many domains. However there are important theoretical and methodological questions which will arise when using deep learning models to gather data for inferential, scientific applications.

### 8.4.2 Computation and Communication

In this section we discuss a number of ways in which computation can improve communication in data analysis[39]. Section 8.4.2.1 below discusses how *literate programming* helps solve many of the issues in reproducibility discussed in Section 8.2.3 above.

#### 8.4.2.1 Literate Programming

*Literate programming* is a concept developed by computer scientist Donald Knuth in the 1980s (Knuth, 1984).

> Literate programming is a methodology that combines a programming language with a documentation language, thereby making programs more robust, more portable, more easily maintained, and arguably more fun to write than programs that are written only in a high-level language. The main idea is to treat a program as a piece of literature, **addressed to human beings rather than to a computer**.

Literate programming is about both cultural and technical developments. Cultural in the sense that it emphasizes the programmer's responsibility to communicate with humans. Lessons learned in English

[38] For example, if google's algorithms mistakenly think someone is dead, then likely the rest of the world will too https://www.nytimes.com/2017/12/16/business/google-thinks-im-dead.html

[39] Our argument is that computation can help communication. Others have taken this idea further and use computation, specifically information theory, as a metaphor for communication e.g. (Doumont, 2009).

courses about organization, clarity, prose, etc are relevant in computer science. Technical in the sense that programming languages need to be altered or created to enable the programmer to write documents which communicate effectively (Knuth, 1984). Literate programming is important to data analysis because data analysis is done with computer code. Explaining the results and details of an analysis involves communicating the details of a computer program.

Traditional data analysis code is not written for human consumption (e.g. little documentation, not made public); the analyst communicates the results through prose and figures which summarize the execution of the code. For some applications this is acceptable in the sense that it accomplishes what is needed. Problems with this style of programming include: it's not reproducible, it's not generalizable, it doesn't communicate the details of the analysis, it can make it harder to find errors with the analysis code, etc. As data analysis becomes more used, more multidisciplinary, and more complex these issues only become more important. The core idea of literate programming – writing code whose target audience includes humans – has a lot to contribute to data analysis.

Communicating data analysis in the medium of code is challenging. Best practices have not yet been established (e.g. how should the code be commented, how to evaluate trade-offs between code clarity and simplicity/efficiency). Software engineers have built up best practices for programming (Wilson et al., 2014, 2017), which can be helpful for data analysts to learn, but may need to be adapted in some cases[40]. In addition to methodology, there are technical developments which make literate programming easier to do.

We highlight *knitr* (Xie, 2015) and R Markdown[41] as examples of research into GDS 6, "science about data science." As discussed in (Donoho, 2017),

> This helps data analysts authoring source documents that blend running R code together with text, and then compiling those documents by running the R code, extracting results from the live computation and inserting them in a high-quality PDF file, HTML web page, or other output product. In effect, the entire workflow of a data analysis is intertwined with

[40] For example, it is often suggested that code comments should describe *why* the code was written the way it was, not *what* the code is doing. For data analysis, where the target audience is probably less experienced with programming, describing the *what* may also be useful.

[41] For more information and examples see `http://rmarkdown.rstudio.com/`.

the interpretation of the results, saving a huge amount of error-prone manual cut and paste moving computational outputs and their place in the document.

There are other examples of technological developments in literate programming for data analysis (e.g. Jupyter notebooks (Perez and Granger, 2015)). Others[42] developing tools to solve technical issues[43] in reproducible research include (Sandve et al., 2013; Kiar et al., 2017). Literate programming is also impactful in statistics education[44], particularly for demonstrating programming examples[45].

#### 8.4.2.2 Open Source

As discussed in Section 8.2.3 reproducibility has a number of technical challenges and literate programming goes a long way to address these issues. Reproducibility crucially demands that the analysis code and data be publically released. This is probably more of a cultural issue than a technical issue (though it's worth pointing out that technologies such as GitHub make sharing code much easier). Some journals (e.g. Biostatistics) encourage the authors to release their code; we hope that more journals will follow this example.

Releasing analysis code makes the original analysis more impactful. A statistics paper is useful primarily to statisticians; an R package is useful to anyone who knows R. If a future analyst has existing code to work with then their life will often become easier. This can take a number of different forms depending on the project.

For some projects, writing an R script which carries out the analysis will allow a future analyst to copy, paste and modify the original script. In other cases, if a researcher develops a new algorithm or complex methodology then releasing a software package may be most appropriate. Resources such as (Wickham, 2015) make developing software packages fairly straightforward. Finally, many research advancements build upon existing algorithms/methodologies. In this case, enhancing an existing software package (as opposed to developing a new one) can be most cost effective. The open source software community has built up best practices for developing and maintaining open source software projects[46].

[42] E.g. see the list of people discussed in https://simplystatistics.org/2015/12/11/instead-of-research-on-reproducibility-just-do-reproducible-research/

[43] The complexity and time costs to making research reproducible is, in part, technical issue.

[44] https://simplystatistics.org/2017/06/13/the-future-of-education-is-plain-text/

[45] E.g. see each of the notes from https://idc9.github.io/stor390/.

[46] E.g. see https://github.com/scipy/scipy/blob/master/HACKING.rst.txt.

It's worth noting there is a large amount of open source statistical software already available, e.g. CRAN, bioconductor, scipy, sklearn. However, we suspect the majority of modern statistics research does not make it into good, publicly available software packages.

Open source extends beyond analysis/research code to educational resources. There are now many technologies which make it easy to publish educational resources for free through a variety of mediums: books, blogs, massive open online courses (MOOCs[47]), websites, etc. As the first author can attest, data science has been particularly good at making resources available for free online. This is an important trend for a number of reasons; it saves students money and provides them with more resources to learn from and it democratizes the subject in the sense that anyone can learn the basic skills to get started in the subject.

The technology is largely available to make most data analyses, statistical algorithms and educational resources freely accessible. The two biggest barriers to releasing these materials are education and incentives. Education in the sense that many data analysts don't currently know these technologies which causes the release of code to be too burdensome. Incentives in the sense that there are few of them to encourage releasing these materials. How would a statistics department hiring committee value a software package vs. a paper? Surely they are both intellectually challenging to produce and valuable, but we suspect the latter would typically receive much more weight that the former. There are some examples of people promoting the value of research software (Sonnenburg et al., 2007) and we hope there will be more in the future.

While open source has lot of potential value, it also raises some questions. For example, not all data can be shared due to privacy/confidentiality reasons; what steps should researchers take in this case to share their analysis code/data? One research avenue worth noting in this regard is *differential privacy* (Smith et al., 2016). As noted in, (Graves et al., 2000; Eick et al., 2001) software decays if it is not actively maintained. This raises the question, for how long should research be reproducible[48]?

### 8.4.3 Education

A number of people have written about updating the statistics curriculum in ways which better reflect the broader definition of data science and the skills required for doing data analysis (Nolan and

[47] http://mooc.org/

[48] https://simplystatistics.org/2017/02/01/reproducible-research-limits/

Temple Lang, 2010; Association et al., 2014; De Veaux et al., 2017; Hardin et al., 2015; Hicks and Irizarry, 2017). A number of programs which have embraced these recommendations have proven to be successful such as the Johns Hopkins Data Science Specialization on Coursera[49] (Kross et al., 2017) and Berkeley's Data8 program[50] (Alivisatos, 2017). We observe three takeaways from this literature (and our own experiences): more computation, more data analysis and the use of open source material.

Communication is another area worth highlighting in this education section because of its importance and ubiquity. The modern data analyst is expected to communicate across a number of different media: written papers/reports, static/dynamic visualizations, online via creating a website/blog, through code, etc. Communication is already under represented in STEM education, in spite of the demand from employers (academic and industrial alike) (Felder and Brent, 2016). Both authors have made effort[51] to include communication in statistics education (Marron, 1999).

#### 8.4.3.1 More Computation

Computation comes into data analysis in a number of ways, from processing data to fitting statistical models to communicating the results. For example, as discussed in Section 8.2.3 many of the issues in reproducibility can be partially addressed by teaching more software engineering to scientists. Many new areas of statistics research involve tackling both statistical and computational issues e.g. see Section 8.4.1.3 on scalable, robust estimators or (Efron and Hastie, 2016).

With the large number of technologies involved with data analysis (programming languages, visualization software, algorithms, etc) one often feels a bit overwhelmed at what one might be expected to know. It is infeasible to know everything. This is where updating the statistics education curriculum is critical. There is probably some rough core set of computational knowledge every trained statistician should have. Given the fixed cost of learning the core computational curriculum, the marginal cost of learning additional computational skills will go down.

[49] https://www.coursera.org/specializations/jhu-data-science

[50] http://data8.org/

[51] E.g. including a lecture on communication in an undergraduate data science course: https://idc9.github.io/stor390/notes/communication/communication.html

#### 8.4.3.2 Pedagogy

The references cited above in Section 8.4.3 make a number of good recommendations for improving statistics education. We highlight a couple of points here.

As many of the above references discuss, the current statistics curriculum often lacks data analysis (Tukey, 1962; Nolan and Temple Lang, 2010). Real data analysis makes the discipline more concrete to students. Focus on solving a real problem can be engaging to students who might otherwise find the subject boring. Teaching data analysis is challenging[52], but it's challenging in the way that teaching the practice of engineering or using the scientific method is challenging. By not giving students practice doing data analysis for a real problem, the statistics curriculum may encourage students to view statistical methodology as a hammer to be procedurally applied to data. It's well established in engineering and the physical sciences that students should get some practical experience doing the thing during their education: why does the same principle not apply more often statistics at the undergraduate level?

When teaching statistical modeling[53] it might be more effective to introduce the model (e.g. linear/logistic regression) in terms of a predictive context instead of the traditional inferential context. While the math behind linear models may not be particularly sophisticated, the concept of randomness and relating it to the real world through data is non-trivial. Moreover, inferential modeling comes with a lot of baggage. For predictive purposes, models such as linear regression can be introduced as an optimization problem which can be heuristically motivated and analytically solved[54]. Once students are comfortable with data modeling we can then introduce inferential/confirmatory modeling.

Finally, we suggest teaching data before statistics in introductory statistics courses; in other words, we should teach exploratory analysis before inferential analysis. This would involve teaching programming, data visualization and manipulation before teaching hypothesis testing. Students are more likely to care about hypothesis testing if they have actually worked on a real problem/dataset which motivates it (as opposed to a hypothetical or artificial one). Berkeley's new course Data 8: The Foundations of Data Science appears to take this approach[55].

[52] https://simplystatistics.org/2017/12/20/thoughts-on-david-donoho-s-fifty-years-of-data-science/

[53] E.g. in an upper level undergraduate course such as UNC's STOR 455: Statistical Methods I.

[54] See for example https://www.inferentialthinking.com/chapters/13/prediction.html.

[55] E.g. see the order of the chapters in the textbook: https://www.inferentialthinking.com/.

## 8.5 Conclusion

So far the arguments in this chapter have been about providing value to society by broadening the discipline in technical ways. Equally as important is increasing diversity in statistics by encouraging women and underrepresented minorities to join and stay in the discipline. Programs, conferences and other efforts such as: ASA Committee on Minorities in Statistics[56], the Women in Machine Learning conference[57] and the Women in Statistics and Data Science conference[58] should be strongly encouraged and expanded upon.

We return to the question of whether data science and statistics are really two different disciplines. If statistics is defined as the narrow discipline described by the quote from John Chambers in Section 8.2 then the answer is yes. However, if statistics embraces the broader idea of greater data science (e.g. by putting more focus on computation in education, research and communication) then we argue the answer is no.

[56] http://community.amstat.org/cmis/home

[57] http://wimlworkshop.org/

[58] http://ww2.amstat.org/meetings/wsds/2018/

## APPENDIX: ADDITIONAL VERTEX CENTRALITY METRICS DETAILS

The code which deals with the Legal Citation network and runs the experiments can be found at: https://github.com/idc9/lawnet/examining_evolution_code.

### Vertex centrality metrics

This Comment focuses on the directed version of the citation network but also considers the undirected version of the network. Most vertex centrality metrics considered are defined for both directed and undirected networks. Ignoring the edge direction means treating citations going into a case the same as citations going out of a case. For example, the degree (undirected) is equal to in-degree plus out-degree. Given the results about out-degree, undirected metrics may be a reasonable choice over directed metrics. This Comment briefly considers the “reversed" network in which the direction of the citations is reversed. This is done primarily to look at a reversed version of PageRank. Some vertex centrality metrics are driven by in-degree (e.g., PageRank, authorities) and other centrality metrics are driven by out-degree (e.g., reversed PageRank, hubs).

Undirected and reversed metrics are considered because of the surprising performance of out-degree. The undirected and reversed metrics tend to perform well, which is further evidence for the importance of out-degree in the evolution of the citation network.

Two time-aware vertex centrality metrics are used: CiteRank and the number of recent citations (Walker et al., 2007). CiteRank is similar to PageRank but down weights older cases. In particular, instead of a uniform “jump" distribution, when CiteRank makes a random jump it selects a new vertex $C$ with probability proportional to $2^{-\frac{\text{age}(C)}{H}}$ exponentially decaying based on case age with half-life = $H$). For the latter metric, the in-degree is computed for each case, but only counting citations that occurred in the most recent $K$ years. The latter is a simple measure of how popular a case is at a given moment in time.

### Sort experiment

This Section discusses some details of the sort experiment. One thousand test cases are selected uniformly at random from all cases between 1900 and 2016, excluding cases that cite zero other cases (i.e., that have zero out-degree).

For each test case, we extract the subnetwork snapshot just before the test case occurs. For example, for a test case on May 15, 1990, we look at the citation network of all cases that occur before May 14, 1990 and call snapshot of the network just before the arrival of the test case. For every case in this network snapshot, we compute each vertex centrality metric we are interested in. Some of the vertex centrality metrics are very computationally intensive, so computing them over and over again takes a long time. We reduce the computational burden by looking at network snapshots once each year from 1900 to 2016 (i.e., look at 116 subnetworks instead of 1000 subnetworks). We then use these annual values to approximate the true values of the centrality metrics at the time of each test case.

The sort experiment compares a ranking of each case with the cases that were actually cited. To compute how well this ranking performed given the actual citations, we use a ranking metric. There are a number of standard ranking metrics we considered: precision, recall, precision at K, and reciprocal rank (Murphy, 2012).

We selected mean rank score, which is defined as follows (Zanin et al., 2009). Suppose we have a ranking of N cases. Suppose $K$ cases are selected and are ranked $R_1, ..., R_k$. The mean rank score is then

$$\frac{1}{K}\sum_{i=1}^{K}\frac{R_i}{N}.$$

The smaller the typical rank, the lower the mean rank score. A random ranking would give a mean rank score value of around 0.5.

Most of the above ranking metrics are used for search engines where one expects the selected results to be near the top of the list. We do not expect a simple vertex centrality metric to place the cited cases near the top of the list. However, we do hope a centrality metrics captures some signal, making the mean rank score more appropriate. We computed all of the above ranking metrics to make sure our results were not sensitive to the particular evaluation choices we made. The results were not qualitatively different for different metrics.

### PageRank time bias

A citation network is a directed acyclic graph (DAG). This Section explains why PageRank is biased to favor older cases in a DAG.

This bias is true because of the way PageRank is defined. One way to describe PageRank is using a random walk around a network. The World Wide Web is a collection of web pages and the links between them. Consider surfing the web for a very long time and jumping from one web page to the next in the following random way. Say you are on webpage X; with probability 0.85, follow one of the links coming from X to one of the webpages X links to; otherwise, with probability 0.15, pick any web page online at uniformly at random.

The PageRank value of a given web page is the proportion of the time the random walk spent at that web page. The intuition is that the more a page is linked to, the more likely the random walk will land on that web page. Furthermore, the more a page X is linked to by pages that are themselves linked to by many pages, the more likely the random walk will land on page X.

For the citation network, PageRank follows citations with a similar random walk. Most random steps follow a citation and go backward in time. This means the random walk will spend more time on older cases.

There are number of other vertex centrality metrics that are driven, at least in part, by out-degree such as hubs or undirected versions of any directed vertex centrality metrics.[59] By ignoring the direction of citations, a citation network can be viewed as an undirected network. In this case metrics such as degree are driven by a combination of both out- and in-degree. Figure 8.5.1 shows that undirected metrics out-perform directed metrics. It is likely this boost in performance comes from the addition of out-degree.

[59] Up until now, the citation network is considered to be a directed graph (i.e., edges go from one case to another case). The citation network can be viewed as an undirected network by ignoring the citation direction.

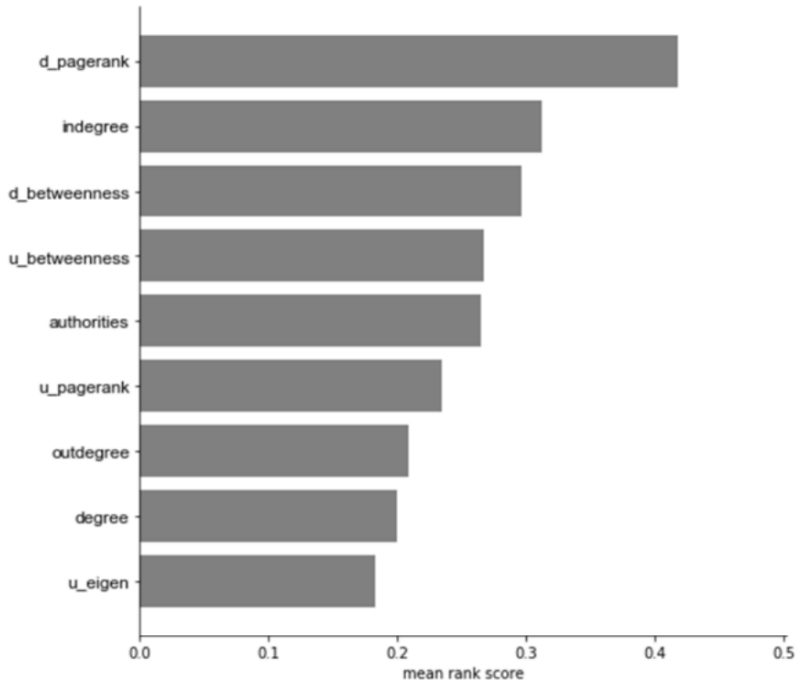



**Figure 8.5.1:** Results of sort experiment for PageRank and Hubs on a reversed graph compared to previous metrics. Hubs performed the best among these metrics and reversed PageRank performed better than all but out-degree and Hubs.

## BIBLIOGRAPHY


Ahn, J., Lee, M. H., and Yoon, Y. J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica*, pages 443–464.

Ahn, J. and Marron, J. S. (2010a). The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259.

Ahn, J. and Marron, J. S. (2010b). The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259.

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

Aldous, D. (1991). Asymptotic fringe distributions for general families of random trees. *The Annals of Applied Probability*, pages 228–266.

Alivisatos, P. (2017). Stem and computer science education: Preparing the 21st century workforce. *Research and Technology Subcommittee House Committee on Science, Space, and Technology*.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07.

Aravkin, A. and Davis, D. (2016). A smart stochastic algorithm for nonconvex optimization with applications to robust machine learning. *arXiv preprint arXiv:1610.01101*.

Arnold, T. and Tilton, L. (2015). *Humanities data in R: exploring networks, geospatial data, images, and text*. Springer.

Association, A. S. et al. (2014). Curriculum guidelines for undergraduate programs in statistical science. *Retrieved March 3, 2009, from http://www. amstat. org/education/curriculumguidelines. cfm*.

Athreya, K. B. and Karlin, S. (1968). Embedding of urn schemes into continuous time markov branching processes and related limit theorems. *Ann. Math. Statist.*, 39(6):1801–1817.

Athreya, K. B. and Ney, P. E. (1972). *Branching processes*. Springer-Verlag, New York-Heidelberg. Die Grundlehren der mathematischen Wissenschaften, Band 196.

Ayat, N.-E., Cheriet, M., and Suen, C. Y. (2005). Automatic model selection for the optimization of svm kernels. *Pattern Recognition*, 38(10):1733–1745.

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric theory*, 13(03):315–352.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1):1–22.

Balch, M. S., Martin, R., and Ferson, S. (2017). Satellite conjunction analysis and the false confidence theorem. *arXiv preprint arXiv:1706.08565*.

Banerjee, S., Bhamidi, S., and Carmichael, I. (2018). Fluctuation bounds for continuous time branching processes and nonparametric change point detection in growing networks. *arXiv preprint arXiv:1808.02439*.

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

Barbero, A., Takeda, A., and López, J. (2015). Geometric intuition and algorithms for ev-svm. *The Journal of Machine Learning Research*, 16(1):323–369.

Bardet, J.-B., Christen, A., and Fontbona, J. (2015). Quantitative exponential bounds for the renewal theorem with spread-out distributions. *arXiv preprint arXiv:1504.06184*.

Barnes, N. (2010). Publish your computer code: it is good enough. *Nature News*, 467(7317):753–753.

Barocas, S., Boyd, D., Friedler, S., and Wallach, H. (2017). Social and technical trade-offs in data science.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bennardo, K. (2014). Testing the geographical proximity hypothesis: An empirical study of citations to nonbinding precedents by indiana appellate courts. *Notre Dame L. Rev. Online*, 90:125.

Bennett, K. P. and Bredensteiner, E. J. (2000). Duality and geometry in svm classifiers. In *ICML*, pages 57–64.

Berger, J. O., Liseo, B., Wolpert, R. L., et al. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28.

Bergeron, F., Flajolet, P., and Salvy, B. (1992). Varieties of increasing trees. In *Colloquium on Trees in Algebra and Programming*, pages 24–48. Springer.

Bhamidi, S. (2007). Universal techniques to analyze preferential attachment trees: Global and local analysis. *preparation. Version August*, 19.

Bhamidi, S., Evans, S. N., and Sen, A. (2012). Spectra of large random trees. *Journal of Theoretical Probability*, 25(3):613–654.

Bhamidi, S., Jin, J., and Nobel, A. (2015). Change point detection in network models: Preferential attachment and long range dependence. *arXiv preprint arXiv:1508.02043*.

Bhardwaj, A. (2017). What is the difference between data science and statistics?

Bickel, P. J. and Levina, E. (2004). Some theory for fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, pages 989–1010.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Black, R. C. and Spriggs, J. F. (2013). The citation and depreciation of us supreme court precedent. *Journal of Empirical Legal Studies*, 10(2):325–358.

Blei, D. M. and Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences*, 114(33):8689–8692.

Boldi, P. and Vigna, S. (2014). Axioms for centrality. *Internet Mathematics*, 10(3-4):222–262.

Bollobás, B. (2001). *Random Graphs*. Number 73. Cambridge University Press.

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

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Bommarito, M. J. (2010). An empirical survey of the population of us tax court written decisions. *Va. Tax Rev.*, 30:523.

Borgs, C., Chayes, J., Daskalakis, C., and Roch, S. (2007). First to market is not everything: an analysis of preferential attachment with fitness. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 135–144. ACM.

Bottou, L., Curtis, F. E., and Nocedal, J. (2016). Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Brodsky, E. and Darkhovsky, B. S. (2013). *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media.

Brown, B. (1983). Statistical uses of the spatial median. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 25–30.

Bryan, K. and Leise, T. (2006). The $25,000,000,000 eigenvector: The linear algebra behind google. *SIAM review*, 48(3):569–581.

Bubeck, S., Devroye, L., and Lugosi, G. (2017). Finding adam in random growing trees. *Random Structures & Algorithms*, 50(2):158–172.

Bubeck, S., Mossel, E., and Rácz, M. Z. (2015). On the influence of the seed graph in the preferential attachment model. *IEEE Transactions on Network Science and Engineering*, 2(1):30–39.

Buckheit, J. B. and Donoho, D. L. (1995). Wavelab and reproducible research. In *Wavelets and statistics*, pages 55–81. Springer.

Bühlmann, P. and Meinshausen, N. (2016). Magging: maximin aggregation for inhomogeneous large-scale data. *Proceedings of the IEEE*, 104(1):126–135.

Bühlmann, P. and Stuart, A. M. (2016). Mathematics, statistics and data science. *EMS Newsletter*, 100:28–30.

Bühlmann, P. and van de Geer, S. (2018). Statistics for big data: A perspective. *Statistics & Probability Letters*.

Carmichael, I. and Marron, J. (2017). Geometric insights into support vector machine behavior using the kkt conditions. *arXiv preprint arXiv:1704.00767*.

Carmichael, I. and Marron, J. (2018). Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*, 1(1):117–138.

Carmichael, I. and Williams, J. (2018). An exposition of the false confidence theorem. *Stat*, 7(1):e201.

Carmichael, I., Wudel, J., Kim, M., and Jushchuk, J. (2017). Examining the evolution of legal precedent through citation network analysis. *NCL Rev.*, 96:227.

Chambers, J. M. (1993). Greater or lesser statistics: a choice for future research. *Statistics and Computing*, 3(4):182–184.

Chapelle, O. and Vapnik, V. (2000). Model selection for support vector machines. In *Advances in neural information processing systems*, pages 230–236.

Chen, P.-H., Lin, C.-J., and Schölkopf, B. (2005). A tutorial on $\nu$-support vector machines. *Applied Stochastic Models in Business and Industry*, 21(2):111–136.

Cherkassky, V. and Ma, Y. (2004). Practical selection of svm parameters and noise estimation for svm regression. *Neural networks*, 17(1):113–126.

Chew, A. and Pryal, K. R. G. (2016). *The Complete Legal Writer*. Carolina Academic Press.

Christmann, A., Luebke, K., Marin-Galiano, M., and Rüping, S. (2005). Determination of hyper-parameters for kernel based classification and regression. *HT014602036*.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1):21–26.

Connecticut, G. v. (1965). 381 us 479 (1965).

Conway, D. (2010). The data science venn diagram.

Craddock, R. C., Jbabdi, S., Yan, C.-G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., Kelly, C., Heberlein, K., Colcombe, S., and Milham, M. P. (2013). Imaging human connectomes at the macroscale. *Nature methods*, 10(6):524.

Crawford, K. (2017). The trouble with bias. Conference on Neural Information Processing Systems, invited speaker.

Crisp, D. J. and Burges, C. J. (2000). A geometric interpretation of v-svm classifiers. In *Advances in neural information processing systems*, pages 244–250.

Cross, F. B. and Spriggs, J. F. (2010). The most important (and best) supreme court options and justices. *Emory LJ*, 60:407.

Csörgö, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*, volume 18. John Wiley & Sons Inc.

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*.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., et al. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, 4:15–30.

Devroye, L. (1998). Branching processes and their applications in the analysis of tree structures and tree algorithms. In *Probabilistic methods for algorithmic discrete mathematics*, pages 249–314. Springer.

Devroye, L. and Lu, J. (1995). The strong convergence of maximal degrees in uniform random recursive trees and dags. *Random Structures & Algorithms*, 7(1):1–14.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2):103–130.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Doumont, J.-l. (2009). Trees, maps, and theorems. *Brussels: Principiae*.

Drmota, M. (2009). *Random trees: an interplay between combinatorics and probability*. Springer Science & Business Media.

Durrett, R. (2007). *Random graph dynamics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*, volume 5. Cambridge University Press.

Eick, S. G., Graves, T. L., Karr, A. F., Marron, J., and Mockus, A. (2001). Does code decay? assessing the evidence from change management data. *IEEE Transactions on Software Engineering*, 27(1):1–12.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*, volume 21. AAAI press Menlo Park.

Felder, R. M. and Brent, R. (2016). *Teaching and learning STEM: A practical guide*. John Wiley & Sons.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 175–185.

Flajolet, P. and Sedgewick, R. (2009). *Analytic combinatorics*. cambridge University press.

Fowler, J. H. and Jeon, S. (2008). The authority of supreme court precedent. *Social networks*, 30(1):16–30.

Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S., and Wahlbeck, P. J. (2007). Network analysis and the law: Measuring the legal importance of precedents at the us supreme court. *Political Analysis*, 15(3):324–346.

Franc, V., Zien, A., and Schölkopf, B. (2011). Support vector machines as probabilistic models. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 665–672.

Fraser, D. A. S., Reid, N., and Lin, W. (2018). When should modes of inference disagree? Some simple but challenging examples. *Annals of Applied Statistics: Special section in memory of Stephen E. Fienberg*.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Garner, B. (2014). *BlackâĂŹs law dictionary*. West Group, 10 edition.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2006). *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science & Business Media.

Gerhardt, M. J. (2011). *The power of precedent*. Oxford University Press.

Gleser, L. J. and Hwang, J. T. (1987). The nonexistence of 100 (1-$\alpha$)% confidence sets of finite expected diameter in errors-in-variables and related models. *The Annals of Statistics*, pages 1351–1362.

Goldschmidt, C. and Martin, J. B. (2005). Random recursive trees and the bolthausen-sznitman coalescent. *Electron. J. Probab*, 10(21):718–745.

Gong, R. and Meng, X.-L. (2017). Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox. *arXiv preprint arXiv:1712.08946*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. book in preparation for mit press. *URL< http://www. deeplearningbook. org*.

Graves, T. L., Karr, A. F., Marron, J., and Siy, H. (2000). Predicting fault incidence using software change history. *IEEE Transactions on software engineering*, 26(7):653–661.

Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., and Giannotti, F. (2018). A survey of methods for explaining black box models. *arXiv preprint arXiv:1802.01933*.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons.

Hand, D. J. et al. (2006). Classifier technology and the illusion of progress. *Statistical science*, 21(1):1–14.

Hannig, J., Iyer, H., Lai, R. C., and Lee, T. C. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515):1346–1361.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., et al. (2015). Data science in statistics curricula: Preparing students to "think with data". *The American Statistician*, 69(4):343–353.

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct):1391–1415.

Henderson, W. D. (2015). What the jobs are: new tech, new client needs create a new field of legal operations. *ABAJ*, 101:36.

Hicks, S. C. and Irizarry, R. A. (2017). A guide to teaching data science. *The American Statistician*, (just-accepted):00–00.

Hitt, M. P. (2016). Measuring precedent in a judicial hierarchy. *Law & Society Review*, 50(1):57–81.

Holmgren, C., Janson, S., et al. (2017). Fringe trees, crump–mode–jagers branching processes and $m$-ary search trees. *Probability Surveys*, 14:53–154.

Hooker, G. and Hooker, C. (2017). Machine learning and the future of realism. *arXiv preprint arXiv:1704.04688*.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer.

Jagers, P. (1975). *Branching processes with biological applications*. Wiley-Interscience [John Wiley & Sons], London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics—Applied Probability and Statistics.

Jagers, P. and Nerman, O. (1984a). The growth and composition of branching populations. *Advances in Applied Probability*, 16(2):221–259.

Jagers, P. and Nerman, O. (1984b). The growth and composition of branching populations. *Adv. in Appl. Probab.*, 16(2):221–259.

Jaggi, M. (2014). An equivalence between the lasso and support vector machines. Technical report, CRC Press.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Janson, S. (2004). Functional limit theorems for multitype branching processes and generalized pólya urns. *Stochastic Processes and their Applications*, 110(2):177–245.

Kiar, G., Bridgeford, E., Chandrashekhar, V., Mhembere, D., Burns, R., Roncal, W. G., and Vogelstein, J. (2017). A comprehensive cloud framework for accurate and reliable human connectome estimation and meganalysis. *bioRxiv*, page 188706.

Knuth, D. E. (1984). Literate programming. *The Computer Journal*, 27(2):97–111.

Kolaczyk, E. (2009). *Statistical Analysis of Network Data*. Springer.

Kolmogoroff, A. (1933). Grundbegriffe der wahrscheinlichkeitsrechnung.

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., and Leek, J. T. (2017). The democratization of data science education. *PeerJ PrePrints*.

Landes, W. M., Lessig, L., and Solimine, M. E. (1998). Judicial influence: A citation analysis of federal courts of appeals judges. *The Journal of Legal Studies*, 27(2):271–332.

Landes, W. M. and Posner, R. A. (1976). Legal precedent: A theoretical and empirical analysis. *The Journal of Law and Economics*, 19(2):249–307.

Lee, M. H., Ahn, J., and Jeon, Y. (2013). Hdlss discrimination with adaptive data piling. *Journal of Computational and Graphical Statistics*, 22(2):433–451.

Leek, J. T. and Peng, R. D. (2015). Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112(6):1645–1646.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and 'sible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM.

Leskovec, J. and Mcauley, J. J. (2012). Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Lohr, S. (2017). Ai is doing legal work. but it won't replace lawyers, yet. *New York Times, March*, 19:2017.

Lu, X., Marron, J., and Haaland, P. (2014). Object-oriented data analysis of cell images. *Journal of the American Statistical Association*, 109(506):548–559.

Lupu, Y. and Fowler, J. H. (2013). Strategic citations to precedent on the us supreme court. *The Journal of Legal Studies*, 42(1):151–186.

Mahmoud, H. (2008). *Pólya urn models*. Chapman and Hall/CRC.

Mariani, M. S., Medo, M., and Zhang, Y.-C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, 10(4):1207–1223.

Maronna, R., Martin, R. D., and Yohai, V. (2006). *Robust statistics*, volume 1. John Wiley & Sons, Chichester. ISBN.

Marron, J. (1999). Effective writing in mathematical statistics. *Statistica neerlandica*, 53(1):68–75.

Marron, J. (2017). Big data in context and robustness against heterogeneity. *Econometrics and Statistics*, 2:73–80.

Marron, J. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753.

Marron, J. S., Todd, M. J., and Ahn, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271.

Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108(501):301–313.

Martin, R. and Liu, C. (2016). Validity and the foundations of statistical inference. *arXiv preprint arXiv:1607.05051*.

Mattera, D. and Haykin, S. (1999). Support vector machines for dynamic reconstruction of a chaotic system. In *Advances in kernel methods*, pages 211–241. MIT Press.

Mavroforakis, M. E. and Theodoridis, S. (2006). A geometric approach to support vector machine (svm) classification. *IEEE transactions on neural networks*, 17(3):671–682.

Members, R.-P. (2017). The r project for statistical computing.

Merryman, J. H. (1953). The authority of authority: What the california supreme court cited in 1950. *Stan. L. Rev.*, 6:613.

Miao, D. (2015). *CLASS-SENSITIVE PRINCIPAL COMPONENTS ANALYSIS*. PhD thesis, University of North Carolina at Chapel Hill.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.

Móri, T. (2007). Degree distribution nearby the origin of a preferential attachment graph. *Electron. Comm. Probab*, 12:276–282.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Naur, P. (1974). Concise survey of computer methods.

Neale, T. (2013). Citation analysis of canadian case law. *J. Open Access L.*, 1:1.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

Nerman, O. (1981). On the convergence of supercritical general (cmj) branching processes. *Probability Theory and Related Fields*, 57(3):365–395.

Network, C. G. A. et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337.

Newman, M. (2010). *Networks: an introduction*. Oxford University Press.

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

Nolan, D. and Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, 64(2):97–107.

Norris, J. R. (1998). *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. Reprint of 1997 original.

Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Patil, D. (2011). *Building data science teams*. " O'Reilly Media, Inc.".

Patil, P., Peng, R. D., and Leek, J. (2016). A statistical definition for reproducibility and replicability. *bioRxiv*, page 066803.

Patty, J. W., Penn, E. M., and Schnakenberg, K. E. (2013). Measuring the latent quality of precedent: Scoring vertices in a network. In *Advances in Political Economy*, pages 249–262. Springer.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060):1226–1227.

Perez, F. and Granger, B. E. (2015). Project jupyter: Computational narratives as the engine of collaborative data science. Technical report, Technical Report. Technical report, Project Jupyter.

Pham, T. (2010). *Some Problems in High Dimensional Data Analysis*. dissertation, University of Melbourne.

Pizer, S. M. and Marron, J. (2017). Object statistics on curved manifolds. *Statistical Shape and Deformation Analysis: Methods, Implementation and Applications*, page 137.

Polson, N. G., Scott, S. L., et al. (2011). Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

Reid, N. (2018). Statistical science in the world of big data. *Statistics & Probability Letters*.

Reid, N., Mukerjee, R., and Fraser, D. (2003). Some aspects of matching priors. *Lecture Notes-Monograph Series*, pages 31–43.

Resnick, S. and Samorodnitsky, G. (2015). Asymptotic normality of degree counts in a preferential attachment model. *arXiv preprint arXiv:1504.07328*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Rowe, V. W. (1973). 410 us 113 (1973).

Rudas, A., Tóth, B., and Valkó, B. (2007). Random trees and general branching processes. *Random Structures & Algorithms*, 31(2):186–202.

Russell, S. and Norvig, P. (2009). Artificial intelligence: A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*.

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS computational biology*, 9(10):e1003285.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. *Neural computation*, 12(5):1207–1245.

Searle, S. R. (1982). Matrix algebra useful for statistics (wiley series in probability and statistics).

Shafer, G. (1976). *A mathematical theory of evidence*, volume 42. Princeton University Press.

Shafer, G. (2008). Non-additive probabilities in the work of Bernoulli and Lambert. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 117–182. Springer.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Smith, M. T., Zwiessele, M., and Lawrence, N. D. (2016). Differentially private gaussian processes. *arXiv preprint arXiv:1606.00720*.

Smythe, R. T. and Mahmoud, H. M. (1995). A survey of recursive trees. *Theory of Probability and Mathematical Statistics*, (51):1–28.

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine learning*, 46(1-3):21–52.

Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., LeCun, Y., MÃžller, K.-R., Pereira, F., Rasmussen, C. E., et al. (2007). The need for open source software in machine learning. *Journal of Machine Learning Research*, 8(Oct):2443–2466.

Spaeth, H., Epstein, L., Ruger, T., Whittington, K., Segal, J., and Martin, A. D. (2014). Supreme court database code book.

Staudte, R. G. and Sheather, S. J. (2011). *Robust estimation and testing*, volume 918. John Wiley & Sons.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science & Engineering*, 14(4):13–17.

Sun, J., Zheng, C., Li, X., and Zhou, Y. (2010). Analysis of the distance between two classes for tuning svm hyperparameters. *IEEE transactions on neural networks*, 21(2):305–318.

Szymański, J. (1987). On a nonuniform random recursive tree. *North-Holland Mathematics Studies*, 144:297–306.

Szymanski, J. (1990). On the maximum degree and the height of a random recursive tree. In *Random graphs*, volume 87, pages 313–324.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Introduction to data mining. 1st.

Tao, T. (2007). What is good mathematics? *Bulletin of the American Mathematical Society*, 44(4):623–634.

Taylor, D., Myers, S. A., Clauset, A., Porter, M. A., and Mucha, P. J. (2017). Eigenvector-based centrality measures for temporal networks. *Multiscale Modeling & Simulation*, 15(1):537–574.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67.

v. Arizona, M. (1966). 384 us 436.

v. Board of Education, B. (1954). 347 us 483.

Van Aelst, S. and Rousseeuw, P. (2009). Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):71–82.

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

Van Der Hofstad, R. (2016). *Random graphs and complex networks*, volume 1. Cambridge university press.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

Walker, D., Xie, H., Yan, K.-K., and Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010.

Wang, H. and Marron, J. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics*, pages 1849–1873.

Wasserman, L. (2014). Rise of the machines. *Past, present, and future of statistical science*, pages 1–12.

Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2-3):149–170.

Wickham, H. (2015). *R packages: organize, test, document, and share your code*. O'Reilly Media, Inc.

Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., Haddock, S. H., Huff, K. D., Mitchell, I. M., Plumbley, M. D., et al. (2014). Best practices for scientific computing. *PLoS biology*, 12(1):e1001745.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., and Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS computational biology*, 13(6):e1005510.

Wu, C. (1998). Statistics = data science? http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf.

Xie, Y. (2015). *Dynamic Documents with R and knitr*, volume 29. CRC Press.

Yao, Y.-C. (1988). Estimating the number of change-points via schwarz'criterion. *Statistics & Probability Letters*, 6(3):181–189.

Yu, B. (2014). Ims presidential address: Let us own data science. http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/.

Yule, G. U. et al. (1925). Ii.âĂŤa mathematical theory of evolution, based on the conclusions of dr. jc willis, fr s. *Phil. Trans. R. Soc. Lond. B*, 213(402-410):21–87.

Zanin, M., Cano, P., Celma, O., and Buldu, J. M. (2009). Preferential attachment, aging and weights in recommendation systems. *International Journal of Bifurcation and Chaos*, 19(02):755–763.

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1):118–132.

Zhang, N. R. and Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.