DISS. ETH NO. 23441

EXTENDING THE HAWKES PROCESS, A GENERAL OUTLIER TEST, & CASE STUDIES IN EXTREME RISK

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH Zürich

(Dr. sc. ETH Zürich)

presented by

Spencer WHEATLEY

MSc Statistics, ETH Zürich

BASc Hon. Engineering/Statistics Option, University of Waterloo

born on 15.09.1988

citizen of Canada

accepted on the recommendation of

Prof. Dr. Didier Sornette,

Prof. Dr Paul Embrechts

2016

Acknowledgements

I specifically acknowledge my supervisor, Professor Didier Sornette, who has been a bottomless source of ideas, wisdom, and positive energy. My PhD work and experience have been deeply enriched, thanks to his supervision. I would also like to thank Professor Paul Embrechts for serving as a co-examiner. I express gratitude to my collaborators, from which I learned a great deal, and the many members of the chair. More generally, I would like to thank the ETH Zürich for employing me at this wonderful institution, as well as the country of Switzerland for its generous funding of scientific research. It deserves mention that my work has also profited greatly from the excellent education in statistics that I have received both at the ETH, and at the University of Waterloo. Importantly, it is due to the support of my family that I have been able to pursue doctoral studies, and thanks to the kindness of friends, who suppressed the natural tendency to ignore statisticians at parties, that the experience was such a pleasure.

Contents

Acknowledgements **1**

Table of Contents **2**

Abstract **6**

I Extensions of the Hawkes Process & the EM algorithm **9**

1 Introduction **10**

2 Hawkes process with Renewal process immigration **19**

2.1 Introduction 19

2.2 The Hawkes Process with Renewal Immigration (RHawkes) 20

2.3 EM Algorithms for the Hawkes Process With Renewal Immigration (RHawkes) 23

2.3.1 The Complete-Data EM Algorithm (EM1) 23

2.3.2 The Semi-Complete-Data EM Algorithm (EM2) 27

2.3.3 Convergence of Hawkes EM Algorithms 28

2.3.4 Computational Efficiency for Estimation of the Hawkes Process 30

2.4 Statistical Inference 31

2.4.1 Likelihood 32

2.4.2 p-Values 33

2.5 Monte Carlo Study of the EM estimation of RHawkes 34

2.5.1 Bias and Efficiency 35

2.5.2 Model Selection 36

2.5.3 Robustness of Branching Ratio Estimation under Misspecification of the Immigration Process . . . 37
2.5.4 Study of EM2 Estimation of Hawkes with Inhomogeneous Poisson Immigration . 40
2.6 Case Study: Self-Excitation of Mid-Price Changes of the E-mini S&P500 Futures . . . 41
2.7 Discussion . . . 43

3 The ARMA point process **45**
3.1 Introduction . . . 45
3.2 The ARMA Point Process . . . 46
3.3 The Relationship to Integer ARMA Models . . . 47
3.3.1 Further details about the INARMA model . . . 50
3.4 Simulation of the ARMA point process . . . 52
3.5 EM algorithm for the estimation of the ARMA point process . . . 54
3.6 Discussion . . . 57

II A general outlier test, & singular "dragon-king" extremes **58**

4 Dragon-kings and extremes **60**

5 A general test for multiple outliers **68**
5.1 Introduction . . . 68
5.2 Outlier Testing Methodology . . . 71
5.2.1 Block, inward, & outward tests . . . 71
5.2.2 Gallery of test statistics . . . 72
5.2.3 EVT & outlier testing in exponential tails . . . 75
5.3 Outlier Test Performance . . . 78
5.3.1 Block test performance compared with a mixture model . . . 78
5.3.2 Masking, swamping, and estimating the number of outliers . . . 79
5.3.3 Comparative study of the performance of sequential estimators . . . 80
5.3.4 Robustness to null mis-specification . . . 82
5.4 Case studies and "Dragon Kings" . . . 85
5.4.1 Drawdowns/crashes in financial markets . . . 86

5.4.2 Nuclear accidents . . . 89
5.4.3 Stock returns . . . 93
5.4.4 Fatalities in Epidemics . . . 96
5.4.5 City sizes . . . 100
5.5 Discussion . . . 101

III Studies of extreme risks **103**

6 Nuclear risk **104**
6.1 Introduction . . . 104
6.2 Data and the measurement of event severity . . . 107
6.2.1 Comparing severity measures & critiquing INES . . . 111
6.2.2 The current fleet of reactors . . . 113
6.3 Uncertainty quantification of risks in nuclear energy systems . . . 113
6.4 Event frequency . . . 115
6.5 Event severity . . . 118
6.6 Runaway disasters as "dragon king" outliers . . . 119
6.7 Modelling aggregate annual damage . . . 122
6.7.1 Quantiles and return periods . . . 122
6.7.2 Expected annual damage . . . 124
6.8 Discussion & policy conclusions . . . 126
6.9 Future Work . . . 129

7 Cyber risk **130**
7.1 Introduction . . . 130
7.2 Data, Results & Methods . . . 133
7.2.1 Data Breach Frequency . . . 134
7.2.2 Data Breach Severity . . . 134
7.2.3 Cumulative Risk & Future Projections . . . 139
7.2.4 Data Breach Risk & Organisation Size . . . 141
7.2.5 Sector & Data Breach Risk . . . 144

7.3 Discussion 145
7.4 Future work 148

Bibliography **153**

Curriculum Vitae **171**

Abstract

This is a cumulative doctoral thesis, which concerns three different areas: I) point process models and their statistical estimation, II) statistical outlier testing, and III) applied statistical studies of risk. These three areas may be, somewhat tenuously, linked to the statistics of extreme events: I) The focal point process models feature contagion/positive-feedback and thus provide a generating process for extreme events. Such models are also used in the modeling and study of extremes; II) Outlier tests are developed, whose generality is due to extreme value theory, and are applied to the detection of singular extremes (so called "dragon-kings" that "live beyond the tail"); III) the two applied studies consider the extreme risk present in accidents in nuclear power generation, and "cyber risk" events where personal information is breached from organizations. More extensive abstracts for these three parts are given below:

I. I consider the *Hawkes process* – which is a *cluster process* and *branching process* – in which *cluster center/immigrant* points follow a Poisson process, and each immigrant may form a cluster of multi-generational offspring. Here, the *Hawkes process* is generalized to have *renewal process* or *Neyman-Scott/shot noise* process immigration. This is named the ARMA (Autoregressive Moving Average) point process, since when aggregated, it is equivalent to the ARMA model for non-negative integer time series. Such generalizations make direct MLE (maximum likelihood estimation) impossible, since one does not know which points are immigrants. EM (Expectation Maximization) algorithms are introduced that enable MLE in such models, improving on the variety of existing estimators, that only "asymptotically" approach MLE performance. Comments are also made on the fast simulation and non-parametric estimation of such models.

II. Next, statistical tests for multiple outliers in exponential samples are considered. Thanks to EVT (Extreme Value Theory), such tests are applicable to general samples, having approximately exponential or Pareto tails. A simple "robust" test statistic is shown to make inward sequential testing – formerly relegated within the literature, since the introduction of outward testing – as powerful as, and potentially less error prone, than outward tests – while being much easier to implement. A comprehensive comparison of test statistics is done, considering performance in both block and sequential tests, and for a variety of null and alternative models. Test sensitivity to misspecification of the sample distribution is studied, and ways to address this such as sample fraction selection and diagnostic methods are discussed. In five case studies significant outliers are detected and related to

the concept of 'Dragon-King' events, defined as meaningful outliers that arise from a unique generating mechanism.

III. Two statistical studies of extreme risks are done, highlighting the important insights about extreme risk that can be obtained: For the risk of nuclear energy systems we provide and analyze a dataset twice the size of the previous best, with a focus on event cost. Comparing cost with the industry standard INES scale demonstrates the inconsistency of INES. Findings include that the rate of accidents dropped significantly after Chernobyl (1986), and has remained roughly constant since. The distribution of costs changed following Three Mile Island (1979) whereby the typical event became smaller, but an extremely heavy tail emerged, being well described by a Pareto distribution with parameter $\alpha = 0.5 - 0.6$. Further significant runaway disasters were found, which we associate with the "dragon-king" phenomenon. It is too soon to evaluate the impact of the industry response to Fukushima. Excluding such improvements, in terms of costs, our range of models suggest that there is presently a 50% chance that a Fukushima event (or larger) occurs every 60-150 years, and that a Three Mile Island event (or larger) occurs every 10-20 years; and that the expected annual cost probably exceeds the cost of a new plant. This highlights the importance of deep improvements to exclude the possibility of future extreme disasters.

For the risk of personal data breaches from organisations, we argue that such events, enabling mass identity fraud, constitute an *extreme risk*. This cyber risk worsens daily as an ever-growing amount of personal data are stored by organisations and on-line, and the *attack surface* surrounding this data becomes larger and harder to secure. Further, breached information is distributed and accumulates in the hands of cyber criminals, thus driving a cumulative erosion of privacy. Statistical modeling of breach data from 2000 through 2015 provides insights into this risk: A current maximum breach size of about 200 million is detected, and is expected to grow by fifty percent over the next five years. The breach sizes are found to be well modeled by an *extremely heavy tailed* truncated Pareto distribution, with tail exponent parameter decreasing linearly, from 0.57 in 2007, to 0.37 in 2015. With this current model, given a breach contains above fifty thousand items, there is a ten percent probability of exceeding ten million. A *size effect* is unearthed where both the frequency and severity of breaches scale with organisation size like $s^{0.6}$. Projections indicate that the total amount of breached information is expected to double from two to four billion items within the next five years, eclipsing the population of users of the Internet.

Abstrakt

Dies ist eine kumulative Dissertation, die sich mit drei verschiedenen thematischen Bereichen befasst: I) Punktprozess Modelle und ihre statistische Bewertung; II) statistische Ausreiertests und; III) angewandte statistische Untersuchungen von Risiken. Verbunden werden diese drei Bereiche mit statistischen Methoden, die der Bewertung extremer Ereignisse dienen: I) Die zentralen Punktprozess Modelle weisen Contagion und positives Feedback auf und stellen dadurch einen Prozess zur Verfgung, der extreme Ereignisse erzeugt. Solche Modelle werde dementsprechend fr die Modellierung und Untersuchung von Extremwerten eingesetzt; II) statistische Ausreiertests, denen die Extremwerttheorie Generalitt verleiht, werden entwickelt und fr die Erkennung einzigartiger extremer Ereignisse eingesetzt (so genannte “Dragon Kings”, die “jenseits heavy-tailed Verteilungen leben”); III) die zwei angewandten Untersuchungen betrachten einerseits die extremen Risiken, die in der nuklearen Energieerzeugung vorhanden sind, anderseits so genannte “Cyber”-Risiken, die entstehen, wenn persnliche Daten Organisationen und Firmen entwendet werden.

Part I

Extensions of the Hawkes Process & the EM algorithm

Chapter 1

Introduction

Point process models are highly interesting in their own right, and enjoy a broad range of application. Here I simply mention some applications in areas of extreme phenomena. For instance, so called *self-exciting* processes and *cluster* processes have been used to model earthquake aftershock dynamics (e.g., [193]), high frequency price fluctuations in financial markets [81] as well as dramatic flash crash events [91], default contagion through business lines [66], insurance claims following a catastrophe [65, 151, 66, 64], clustering of extreme returns for risk measure estimation [47], and even the spread of violence [165] and crime [163]. Further, *branching processes* [114], which may be related to point processes, were used to model nuclear chain reactions in the Manhattan project.

Here the necessary terminology of univariate point process models is given, with an emphasis on points occurring in time. For a basic reference see [60], and for a more advanced reference see [62]. Consider a sequence of random event times $\{T_i\}_{i\in\mathbb{Z}}$, such that $T_i < T_{i+1}$ with *inter-event (waiting) times* $W_i = T_i - T_{i-1}$. This sequence $\{T_i\}$ defines an *univariate point process* with *counting process* $N(t) = \sum_i 1_{T_i \le t}$. Denote a realization of the point process by $\boldsymbol{t}_{1:n} = \{t_1, \ldots, t_n\}$, on window $(0, r]$, where $N(r) = n$, and the *history* of the process is captured by the *natural filtration*, $\mathcal{F}_t = \{t_i : t_i < t\}$ where $i \in \mathbb{Z}$, i.e., including events on the negative part of the time axis as well. In general, the abbreviation $i : j = i, i+1, \ldots, j-1, j$ may be used for such sets of indices. Note that a point process is *marked* if an associated variable, e.g., a size, is associated with each point. For simplicity the notation here will exclude marked processes, with little loss of generality.

A point process can be defined by its conditional intensity function (CIF),

$$\lambda(t|\mathcal{F}_t) = \lim_{\Delta\to 0^+} \Delta^{-1}\mathrm{E}\left[N(t,t+\Delta)|\mathcal{F}_t\right] \ . \tag{1.1}$$

For instance, consider a *renewal process*, which has waiting times W_i, say with $\Pr\{W_i \in (0,\infty)\} = 1$, that are i.i.d realisations of a random variable with pdf (probability density function) $g(\cdot)$. In this case the conditional intensity $\mu(\cdot)$, and pdf $g(\cdot)$, are related via equations,

$$\mu(w) = \frac{g(w)}{1-\int_0^w g(s)ds}, \quad g(w) = \mu(w)\exp\left(-\int_0^w \mu(s)ds\right). \tag{1.2}$$

A constant intensity $\mu > 0$ uniquely corresponds to an exponential density (a Poisson process). Any non-exponential waiting time pdf will have a non-constant renewal intensity $\mu(w)$. Thus, the Poisson process is the "memoryless" point process, that evolves "without aftereffects" [146]. Thus, the Poisson process may be considered as a null model, against which one my try to detect clustering or other features [193].

A useful class of processes are *cluster processes*, where one process defines so-called *cluster center* points, and then, for each cluster center, associated cluster points are introduced by a separate, and typically independent, clustering mechanism. These cluster processes may also be thought of as *branching processes* [114] where the cluster center points are *immigrants*, and some *parent* points are able to produce *offspring* points via *triggering functions*. All of the parent-offspring relationships are captured in the *branching structure*. This can also be thought of as a *causal structure* (i.e., a directed graph) [78]. It is almost always the case that the cluster center/immigrant process is a Poisson process, as the memoryless property substantially symplifies analysis (and estimation) [60]. Regarding cluster formation, in a *NS (Neyman-Scott)* process, for each immigrant, a single burst of a random number of offspring is generated, independently of both their associated immigrant, as well as the other clusters. I.e., the NS clusters have a single generation of offspring. From here onwards, a Poisson distribution for the number of offspring will be considered. This NS process is also a *shot-noise process*. Such processes have been used for modeling earthquake aftershocks [281], rainfall events [59, 213], insurance claims following a catastrophe [65, 151, 66, 64], etc. Since the immigrants are the only fertile points within the NS process, to determine the CIF one needs to know which points are immigrants. For this, one uses the indicator variable $Z_i^\mu = 1$ if t_i is an immigrant and $Z_i^\mu = 0$ otherwise. That is, $Z_i^\mu = dN^\mu(t_i)$

where $N^\mu(s) := \int_0^s dN^\mu(s)ds$ is the immigrant counting process. Thus the NS process has CIF,

$$\lambda(t|\mathcal{F}_t^Z) = \mu + \int_{-\infty}^{t} \theta(t-s)dN^\mu(s) \ , \tag{1.3}$$

where cluster center points are included in the realisation, and the immigrant indicator variables are contained within the filtration $\mathcal{F}_t^Z$. The constant $\mu \in (0, \infty)$ is the intensity of Poisson immigration, and $\theta(s) = \gamma g(s)$ is the *triggering/self-excitation* function, and is the IPP (Inhomogeneous Poisson Process) originating at each immigrant. Thus it may also be called the *offspring intensity*. The function $g(\cdot)$ is the *offspring density*, with $\int_{-\infty}^0 g(s)ds = 0$, providing the law for the waiting time between parents and offspring. The parameter $\gamma \geq 0$ is the mass of $\theta(\cdot)$, and is a *branching ratio*, i.e., the expected number of offspring of each point. In this case, including the immigrant point, the expected cluster size is $\gamma + 1$, and the model is stationary for finite γ.

An important cluster process is the *Hawkes process* [116, 119], in which immigrants follow a Poisson process, and then multiple generations of offspring are triggered identically, relative to their parent. Thus each immigrant, along with its multi-generational tree of offspring, form a *cluster*. Extensions of this model have become the predominant model for earthquake aftershocks (ETAS) [193, 121], financial price fluctuations (e.g., [47, 35, 91, 22]), as well as many other applications. The Hawkes process has also long been referred to as the "autoregressive" point process [60]. The Hawkes CIF is,

$$\lambda(t|\mathcal{F}_t) = \mu + \Phi(t|\mathcal{F}_t) \ , \ \Phi(t|\mathcal{F}_t) = \sum_{i:t_i<t} \eta f(t-t_i), \tag{1.4}$$

where $\mu \in (0, \infty)$ is the deterministic (and possibly varying) *immigration intensity*, and the *triggering/self-excitation* functions $\phi(\cdot) = \eta f(\cdot)$, also called the *offspring intensities*, are IPP originating at each observed point, t_i. The *offspring density* $f(\cdot)$, with $\int_{-\infty}^0 f(s)ds = 0$, again provides the law for the waiting time between parents and offspring. Here the branching ratio is $\eta \geq 0$. Due to the autoregressive nature of the process the expected cluster size is $\frac{1}{1-\eta}$, and the model becomes non-stationary for $\eta > 1$ (the critical case $\eta = 1$ is borderline stationary with non-standard scaling properties [217]). A realization of this CIF (1.4) is shown in the top panel of Figure 1.1, allowing one to observe the multi-generational branching structure. It is worth noting that the CIF (1.4) is just a superposition of the immigration and the independent IPP that each point generates. Unlike in the NS process, here one does not need to know which points are immigrants.

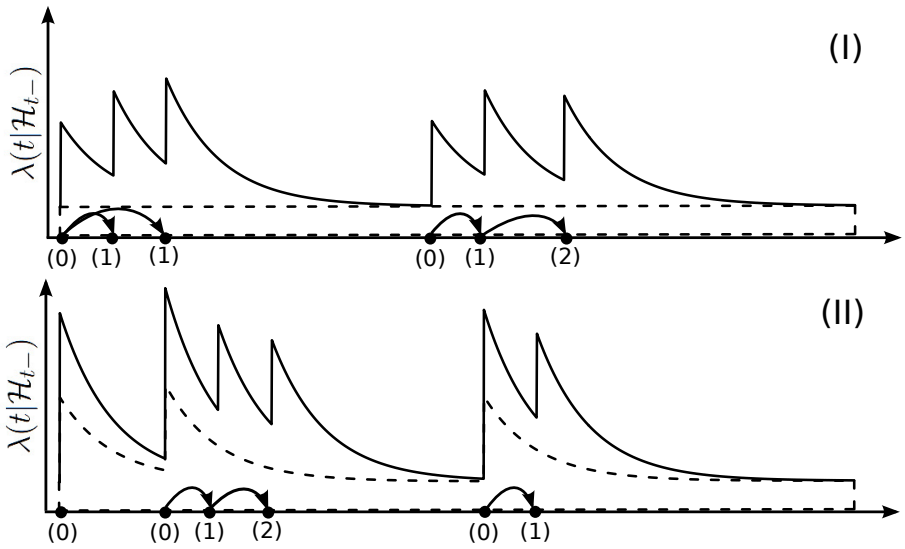

Figure 1.1: Illustration of the CIFs for (I) the Hawkes process (1.4) and (II) the Hawkes process with over-dispersed renewal process immigration (2.1). The dotted line represents the intensity of immigration, and the solid line corresponds to the total intensity. The arrows indicate parenthood and the numbers in parentheses indicate the generation of offspring.

Aggregation

Quoting Brillinger [37]: "many discrete variate processes arise as aggregated point processes, i.e., counts of a point process in consecutive intervals of time (bins). Models, techniques, and theoretical results developed for the study of time series and point processes can suggest corresponding techniques for each other." Also, the Hawkes process has also long been referred to as the "autoregressive" point process [60]. In [150] it was shown that the integer valued autoregressive model (INAR) [181, 285, 97] approximates the Hawkes process. Similarily, a Hawkes process aggregated on a grid is approximately an INAR process. Further, as the grid becomes sufficiently fine, and an infinite order INAR model is considered, the models converge (weakly) [150]. With the same reasoning, it is clear that the NS process (1.3), when aggregated, forms an integer valued moving average (INMA) process. This is discussed further in Chapter 3.

Second order statistics

An important object is the *autocovariance density*, $c(u)dudt = \mathrm{Cov}(dN_t, dN_{t+u})$, where $N(t) = \int_0^t dN(s)ds$, and $c(u)dudt$ can be thought of as the covariance of the counting measures $N(du)$, $N(dt)$ on windows du and dt, separated by u [60]. Defining *palm intensity* $h(u) = \mathrm{P}\{dN_{t+u} = 1|dN_t = 1\}$ for $u > 0$, as the intensity conditioned only on a previous event, and *unconditional intensity* $\bar{\lambda} = \mathrm{P}\{dN_t =$

1}, then the second moment is $\mathrm{E}[dN_t dN_{t+u}] = \bar{\lambda} h(u)$ for $u > 0$. Thus, for a stationary point process the *autocovariance density* is,

$$c(u) = \bar{\lambda}\delta(u) + \bar{\lambda}h(u) - \bar{\lambda}^2 \ , \tag{1.5}$$

where $\delta(u)$ is the standard dirac function. Thus, the palm intensity defines the autocovariance of the process. The palm intensity is also easy to estimate in practice, e.g., with a histogram type estimator. The Fourier transform of the covariance density is the *spectrum*, which may be studied as in stationary time series. For the Hawkes process the palm intensity is an integral equation,

$$h_{AR}(u) = \mu + \phi(u) + \int_0^\infty \phi(s) h_{AR}(u-s) ds \ , \tag{1.6}$$

which may be generally solved by Wiener-Hopf methods [60]. For the NS process (1.3), the palm intensity is,

$$h_{MA}(u) = \mu(1+\gamma) + \frac{\gamma(2+\gamma)}{1+\gamma} \int_0^\infty g(s) g(s+u) ds. \tag{1.7}$$

When ϕ is exponential, then (1.6) can be solved by Laplace transform ([60] page 67),

$$h_{AR}^{Exp}(u) = \frac{\mu}{1-\eta} + \frac{\eta}{2}\frac{(2-\eta)}{2(1-\eta)}\frac{1}{\beta}\exp\{-u/(\beta/(1-\eta))\} = h_{MA}^{Exp}(u; \gamma = \eta/(1-\eta), \tilde{\beta} = \beta/(1-\eta)) \tag{1.8}$$

Thus the exponential kernel provides an example of when the second order statistics of the Hawkes and NS model are the same. That is, the palm intensity for the Hawkes process renormalizes the parameters of the Hawkes kernel by equating the expected number of cluster offspring $\eta = \frac{\gamma}{1+\gamma}$ and cluster characteristic length $\tilde{\beta} = \beta/(1-\eta)$. Considering the analogy with the integer time series, such representations relate to the AR and MA representations of ARMA time series [39]. The depth to which these representations are equivalent is not explored here, however it becomes relevant when one performs estimation based on the palm intensity h rather than the full conditional intensity (see Sec 1). For instance, from the second order statistics (e.g., the palm intensity) one cannot distinguish between the Hawkes process and NS process with exponential offspring densities (1.8). On a related note, in [150] it was shown that the INAR model can be represented by both AR and MA forms of *standard* (i.e., real-valued) time series.

Simulation

Given the conditional intensity (1.1) of a point process, one can simulate the model directly via a thinning algorithm [166]. For a Hawkes process one could do this by simulating a single point, updating the conditional intensity, and then repeating. This method continues to be implemented [115]. Other brute force methods (e.g., [197]) have been discarded in favour of the current preferred method [184, 192] which decomposes the simulation into independent immigrant and offspring parts. This method exploits the branching process formulation. That is, each realized point t_i generates the IPP offspring process, $\phi(t - t_i) = \eta f(t - t_i)$, from which one simulates. In the proposed methodology, as well as applications, thinning has also been used for simulating these IPP. For the typical case of pdf f which decays – especially when the decay is strong, such as in a (shifted) Pareto pdf, which is employed in seismology [278] – this thinning is highly inefficient. Due to this a fast algorithm for the special case of the exponential pdf was developed in [67], exploiting the Markovian structure of that specific parameterisation.

To be clear, the *thinning method* of [166] requires generating a *proposal process* with intensity that exceeds the intensity of the model from which one wants to draw a sample (the *proposal process*). Then one retains/accepts each proposal point with probability equal to the ratio of the target intensity to that of the proposal intensity, at that point. Analogously to the accept/reject sampling from a density, using thinning is justified when inverse transform sampling is not possible. The cost of this generality is inefficiency. The *inversion method* [51], which is equivalent to inverse transform sampling from an interevent time pdf, relies on the fact that $T_1, T_2, \dots$ is an IPP with *compensator* $\Lambda(t) = \int_0^t \lambda(s)ds$, if and only if $\Lambda(T_1), \Lambda(T_2), \dots$ is a unit Poisson process. Thus, one can simulate from a unit Poisson process and transform the realization with $\Lambda^{-1}(t)$, to obtain a sample from the IPP with perfect efficiency.

In other words, and a slightly different formulation: The result of simulating from an IPP, e.g., $\phi(\cdot) = \eta f(\cdot)$ from the Hawkes process (1.4), is a sample of $N_\phi \sim \text{Pois}(\eta)$ inter-event times that are i.i.d from pdf f. Thus, rather than thinning, one can instead sample the number of offspring N_ϕ, and then sample the distance of each offspring from their parent i.i.d. with pdf f by *inverse transform*: $X \overset{d}{=} F^{-1}(U),\ U \sim \text{UNIF}(0, 1)$. Therefore all that is required is that the inverse CDF F^{-1} be known.

Estimation

Given a realization $\boldsymbol{t}_{1:n}$, on window $(0, r]$, with $N(r) = n$, the likelihood of a point process, with parameter vector $\boldsymbol{\theta}$, is,

$$\mathrm{L}(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}) = f(\boldsymbol{t}_{1:n};\boldsymbol{\theta})\mathrm{Pr}_{\boldsymbol{\theta}}\left[N(r) - N(t_n) = 0|\boldsymbol{t}_{1:n}\right]. \tag{1.9}$$

This is the product of the *joint density*, and the probability that no events occur between the last point t_n and the end time time r [62]. This joint density can be factored into a product of conditional marginal densities,

$$\mathrm{f}(t|\boldsymbol{t}_{1:N(t)}) = \lambda(t|\boldsymbol{t}_{1:N(t)})\exp\left(-\int_{t_{N(t)}}^{t} \lambda(s|\boldsymbol{t}_{1:N(s)})ds\right), \tag{1.10}$$

each of which is the conditional probability of observing a point at t times the probability of no points between the previous point $t_{N(t)}$ and t. For an IPP, the density (1.10) becomes unconditional; and for a homogeneous Poisson process it becomes an exponential density. Thus, the log-likelihood is,

$$\log \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}) = \sum_{i=1}^{n} \log \lambda(t_i|\boldsymbol{t}_{1:N(t_i)}) - \int_0^r \lambda(s|\boldsymbol{t}_{1:N(s)})ds \tag{1.11}$$

Here the conditional intensity is only conditional on the observed points rather than the full history of the process, contained within the filtration [191]. Maximization of this log-likelihood (1.11) with respect to $\boldsymbol{\theta}$ maximizes the intensity at observed points while minimizing it where no points are observed.

If the conditional intensity cannot be evaluated – such as for the NS model, when the immigrant points are unknown, as well as its aggregated INMA form – then likelihood maximization cannot be done. This happens in time series models with an MA component since the innovations are unknown, and in the point process case where, e.g., the immigration is not Poissonian (2.1), because the branching structure is not observed. For instance, in [36] the INMA process is estimated by conditional least squares or generalized method of moments. For NS processes one often uses method of moments, or maximizes an quasi-likelihood based on the palm intensity [272, 205]. This "palm likelihood" technique exploits the fact that the palm intensity (1.5) is easy to estimate, while the full conditional intensity is not. This approach has also been taken for Hawkes processes [21], although it is not necessary to do so. It also introduces the issue of being able to identify/discriminate between e.g., the Hawkes process and the NS process, which can have equivalent second order statistics (1.8). In [65] NS/shot-

noise processes have been estimated by filtering the (unobserved) intensity under a Gaussian process approximation. For the INARMA models, a Bayesian MCMC technique was used [188, 83]. Such quasi MLE approaches have also gone to the frequency domain [38] for the Hawkes and NS processes [118], NS processes with renewal immigration [209, 46] and INAR models [231]. Of course these methods are only asymptotically competitive with MLE. Instead, one can often use an EM algorithm to perform MLE, which is developed here.

There has been some interest in the nonparametric estimation of Hawkes process models. In particular, the INAR approximation of the Hawkes process has been used to estimate a discretized Hawkes process [149]. Further, the estimators in [22, 21] can be made non-parametric: given any estimate of the palm intensity (e.g., a histogram), one solves for Hawkes kernel ϕ via the Wiener-Hopf integral equation (1.6). For general parametric models, and non-parametric models, the equation must be solved with a numerical quadrature scheme. In [279, 163] it was shown that EM algorithms (Sec. 1) are natural for the estimation of Hawkes processes, and that nonparametric estimation is easy to implement. In [185] it was shown that nonparametric implementations of the EM algorithm were superior to the INAR approximation method [149] and the solution of the Wiener-Hopf equation (1.6) of [22, 21]. In particular, the EM algorithm naturally allows one to use non-parametric estimators that are more efficient than histogram estimators, for instance spline estimators with adaptive bandwidth/smoothness [154]. The non-EM methods, on the other hand, require perhaps impractically large sample sizes for precise estimation – especially in the tails – as is seen in the simulation studies [22, 21, 185].

EM algorithm

An Expectation Maximization (EM) algorithm [71] is an iterative algorithm to obtain the MLE parameters of a model, where *observed data* $\boldsymbol{X}$ is known, but the model depends on some latent or *missing data* $\boldsymbol{Z}$. In such a case it may be difficult to perform MLE with the *incomplete data likelihood* $L(\boldsymbol{\theta}|\boldsymbol{X})$, but easier with the *(complete data) likelihood* $L(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Z})$. Since $\boldsymbol{Z}$ is unknown, it is accounted for probabilistically.

In an EM algorithm, one first gives an initial parameter estimate $\hat{\boldsymbol{\theta}}^{[0]}$. Then each m^{th} iteration consists of two steps:

1) Given the estimates $\hat{\boldsymbol{\theta}}^{[m]}$, in the *expectation step* (E-step), one needs to calculate the function

$$\mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}^{[m]}) = \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}^{[m]}}\Big[\log \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Z})\Big], \tag{1.12}$$

which is the expected value of the log-likelihood with respect to the conditional distribution of missing data $\boldsymbol{Z}$, given the observed data $\boldsymbol{X}$ and current estimates $\hat{\boldsymbol{\theta}}^{[m]}$.

2) The *maximization step* (M-step) consists in determining the solution of

$$\hat{\boldsymbol{\theta}}^{[m+1]} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}^{[m]}), \tag{1.13}$$

that is, the expected likelihood (1.12) is maximized to obtain new estimates $\hat{\boldsymbol{\theta}}^{[m+1]}$. The algorithm iterates the E and M steps until the parameter estimates $\hat{\boldsymbol{\theta}}^{[m]}$ stabilize. With each iteration, parameter estimates are guaranteed not worsen the observed data likelihood worse [71].

In general, for point processes with a branching process interpretation – such as the Hawkes and NS processes – an EM algorithm can be developed where the missing data is the branching structure [279, 163]. Given the branching structure, the process is then decoupled into identical (up to a time shift) and independent IPP. Then the statistical problem in the M step is density estimation with sample weights. This will be further illustrated in the subsequent chapters, where the Hawkes process is extended to have renewal process immigration, and in another case, to have NS/shot noise process immigration. In both cases, evaluating the CIF requires knowing which points are immigrants, and thus MLE is made possible through treating this as missing data and using the EM algorithm.

Chapter 2

Hawkes process with Renewal process immigration

This chapter is based on the paper [287].

2.1 Introduction

In the Hawkes processes, the immigration, i.e., the process defining the location of clusters, is a Poisson process. In this form, including the case of inhomogeneous Poisson immigration, the Hawkes model can be estimated by Maximum Likelihood Estimation (MLE). We introduce a natural extension: the *Hawkes process with renewal process immigration* (RHawkes). The renewal process, like the Poisson process, has i.i.d (independent and identically distributed) waiting times, but with an arbitrary waiting time distribution rather than the exponential distribution of the Poisson process. For instance, with a Weibull waiting time distribution, immigration ranges from being highly dispersed to having highly regular spacing. As we will see in the case-study in the present paper, such an extension can provide superior quality of fit in applications. However this flexibility comes at the cost of making direct MLE practically impossible.

To the best of our knowledge, the RHawkes model has not been considered in the literature. There have been some similar models proposed for applications in climatology but with either very restrictive model assumptions, or with less desirable estimation properties: In [58], a renewal cluster model was proposed for clustering of rainfall events. This model featured Bartlett-Lewis type clustering [60], where offspring are distributed after their immigrant in a finite renewal process with random termination size.

In this specification, no overlap of clusters is allowed. This severe simplification allows for easy MLE. In [219], a Bartlett-Lewis cluster process with renewal process immigration was considered for clustering of rainfall events to better account for the occasional observation of long periods without rainfall. The authors estimated the model via maximization of a quasi-likelihood. In [46] a renewal NS process was proposed with a spectral estimation method used.

Here, we propose an *Expectation Maximization* (EM) algorithm [71] for estimation of the RHawkes model. An EM algorithm for the (standard) Hawkes process (here called EM0) has already been developed [178, 279, 164]. We have extended this approach to the case of renewal immigration (EM1), and further introduced another EM algorithm with a reduced set of missing data (EM2). This second algorithm also allows for the computationally efficient estimation of the Hawkes process with IPP immigration. Both EM algorithms may be easily extended to multivariate Hawkes models, its' spatio-temporal extensions, as well as marked processes, but for simplicity of presentation we will focus on the basic model.

The structure of the chapter is as follows: Section 2 presents the RHawkes process. Sections 3 introduces the EM1 and EM2 algorithms for Hawkes and RHawkes models. In section 4, the computation of the likelihood and goodness of fit tests for RHawkes are discussed. Section 5 presents Monte Carlo studies on: the consistency of EM1 estimation of RHawkes, model selection, and robustness of branching ratio estimation under immigrant process misspecification. In section 6, a case study is done on the estimation of Hawkes and RHawkes on high frequency price changes in financial markets. In section 7, we conclude with a discussion.

2.2 The Hawkes Process with Renewal Immigration (RHawkes)

We propose to extend the Hawkes process by considering renewal process immigration (1.2). Thus the immigration process will renew at each immigrant point, which may be identified within the realisation by $Z_i^\mu = 1$ if t_i is an immigrant and $Z_i^\mu = 0$ otherwise. That is, $Z_i^\mu = dN^\mu(t_i)$ where $N^\mu(s) := \int_0^s dN^\mu(s)ds$ is the immigrant counting process. Thus the CIF of the *Hawkes Process with renewal immigration* (RHawkes) is given by,

$$\lambda(t|\mathcal{F}_t^Z) = \mu(t - t_{I[N(t)]}) + \Phi(t|\mathcal{F}_t), \tag{2.1}$$

where the full history including immigrant indicator variables are included in the filtration $\mathcal{F}_t^Z$, and $\mathcal{F}_t$ omits the immigrant indicator variables. The self-exciting part, $\Phi(t|\mathcal{F}_t)$, is the same as in (1.4), and the index function,

$$I[N(t)] = \max(j \in \{1, \ldots, N(t)\} \ : \ Z_j = 1) \ , \tag{2.2}$$

returns the index of the most recent immigrant point prior to time t. A realization of the RHawkes process (2.1) together with the immigrant intensity is presented in the lower panel of Figure 1.1. This substantial generalization allows for dependence between clusters. Indeed, in the Hawkes process (and *Poisson cluster processes*), different clusters are independent. In contrast, RHawkes is a *renewal cluster process*, featuring "nearest neighbour" dependence between clusters. Further extensions can follow by allowing more extensive interaction of clusters. An instance of this would be to let the cluster locations themselves follow a Hawkes process.

For the Hawkes model, the CIF (1.4) can be evaluated and thus MLE can be performed, whereas for RHawkes one needs to know which events are immigrants to evaluate the likelihood (1.9), and thus, in most practical cases, direct MLE is not possible. In the following section, the later problem is formulated as a missing data problem with the EM framework to enable estimation. As will be seen, the EM1 and EM2 algorithms allow for easy estimation of the RHawkes model as long as the immigrant intensity $\mu(\cdot)$ and self-exciting intensity $\Phi(\cdot)$ do not have common parameters.

Suppose the immigrants of the process are known: define $z_i = 1$ if t_i is an immigrant and $z_i = 0$ if not. Given these point types (the so-called *immigrant vector* $\boldsymbol{z}_{1:n}$), the log-likelihood of a realization can be written in the form,

$$\begin{aligned}\log \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{z}_{1:n}) = \sum_{i=1}^{n} 1_{\{z_i=1\}} \log(\mu(t_i - t_{I[i]})) - \int_0^r \mu(s - t_{I[N(s)]}) ds \\ + \sum_{i=1}^{n} 1_{\{z_i=0\}} \Phi(t_i|\boldsymbol{t}_{1:N(t_i)}) - \int_0^r \Phi(s|\boldsymbol{t}_{1:N(s)}) ds, \end{aligned} \tag{2.3}$$

which has two separate parts: the first summand and integral for the immigrant renewal process, and last two for the clustering/offspring part. [191] provides the regularity conditions and asymptotic properties of MLE for such point process, with renewal and Hawkes processes as examples. Further, MLE of a renewal process is nothing more than density MLE from i.i.d samples, and thus standard regularity conditions and asymptotic properties apply. For the clustering part of the log likelihood (2.3)

to exist, the upper endpoint of the offspring density must be greater than the largest observed inter-event time. Aside from this, when the regularity conditions of [191] are satisfied for the renewal process and the Hawkes process (for instance, densities are not too heavy-tailed), they will also be satisfied in (2.3).

When the immigrant vector is unknown and treated as random, the log likelihood should be defined via an expectation of the conditional likelihoods (2.3) with proper probabilities for all possible immigration vectors:

$$\log \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}) = \log \left(\sum_{j=1}^{2^{n-1}} \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{1:n}^{(j)}) \Pr[\boldsymbol{Z}_{1:n}^{(j)}|\boldsymbol{\theta}] \right). \tag{2.4}$$

The weighting probabilities depend on the true model parameters, thus direct maximization of (2.4) is not possible. Instead, an iterative Expectation Maximization (EM) algorithm shall be used.

Suppose now that we consider the full branching structure $\boldsymbol{Z}_{n\times n}$ as missing data. The branching structure is represented by a lower-triangular matrix $\boldsymbol{Z}_{n\times n}$ with diagonal elements $Z_{i,i} = 1$ if point t_i is an immigrant, and $Z_{i,i} = 0$ if not; and sub-diagonal elements $Z_{i,j} = 1,\ j < i$, if point t_j is parent to point t_i. Since a point can be either an immigrant or an offspring of a single parent, each row of the matrix has one unit element.

Following (1.9), the complete likelihood $L(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{n\times n})$ is constructed as a product of the joint pdf of observed events $f(\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{n\times n}) = f(\boldsymbol{t}_{1:n}|\boldsymbol{Z}_{n\times n}) f(\boldsymbol{Z}_{n\times n})$ and another term (the compensator) which accounts for the probability of observing no event after the last event in each independent subprocess. Thus, after substituting (2.7) into (1.9) and rearranging, we can write the complete log-likelihood of RHawkes:

$$\begin{aligned} \log \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{n\times n}) = \log f(\boldsymbol{Z}_{n\times n}) &+ \left[\sum_{i=1}^{n} \sum_{j=J[i]}^{i-1} Z_{i,j} \log \eta f(t_i - t_j) - \int_0^r \Phi(s|\boldsymbol{t}_{1:N(s)}) ds \right] \\ &+ \left[\sum_{i=1}^{n} \sum_{j=1}^{i-1} Z_{i,i} 1_{\{I[i]=j\}} \log \mu(t_i - t_j) - \sum_{i=1}^{n+1} \sum_{j=1}^{i-1} 1_{\{I[i]=j\}} \int_{t_j}^{t_i} \mu(s - t_j) ds \right], \end{aligned} \tag{2.5}$$

where we denoted $t_0 = 0$ as the starting time and $t_{n+1} = r$ as the stopping time. Neither of these points are included in the sample. Similarly, if we consider only the semi-complete data (i.e., missing data being $Z_i = 1$ if t_i is an immigrant and $Z_i = 0$ otherwise) the *semi-complete log-likelihood* (2.6) is

given by,

$$\log \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{1:n}) \propto \sum_{i=1}^{n}(1-Z_i)\log\Phi(t_i|\boldsymbol{t}_{1:N(t_i)}) - \int_0^r \Phi(s|\boldsymbol{t}_{1:N(s)})ds + \sum_{i=1}^{n}\sum_{j=1}^{i-1} Z_i 1_{\{I[i]=j\}}\log\mu(t_i-t_j) - \sum_{i=1}^{n+1}\sum_{j=1}^{i-1} 1_{\{I[i]=j\}}\int_{t_j}^{t_i}\mu(s-t_j)ds, \quad (2.6)$$

where $P[Z_i = 1] = 1$, if $\Phi(t_i|\boldsymbol{t}_{1:N(t_i)}) = 0$.

2.3 EM Algorithms for the Hawkes Process With Renewal Immigration (RHawkes)

An EM algorithm for the (standard) Hawkes process (EM0) was identified in [178] and formalized in [279, 163]. It allows for both parametric and non-parametric estimation of the Hawkes process, including with IPP immigration. In the following subsections, we introduce algorithm EM1, which extends EM0 to the RHawkes case, and EM2 which has a reduced definition of missing data. In the same way that EM2 simplifies EM1, EM0 may be simplified, in which case we call this EM2 for the Hawkes process. A comparison of the computational efficiency of these algorithms and the theory for convergence follow.

2.3.1 The Complete-Data EM Algorithm (EM1)

For the *complete-data EM algorithm* for RHawkes (EM1), the observed data $\boldsymbol{X}$ is the point realization $\boldsymbol{t}_{1:n}$, and the unobserved data $\boldsymbol{Z}$ is the *branching structure* of the process which indicates: (i) immigrant events and (ii) parenthood of offspring events (see Figure 1.1).

The main step in deriving (1.12) for RHawkes (2.1) is the conditional density,

$$f(\boldsymbol{t}_{1:n}|\boldsymbol{Z}_{n\times n}) = \prod_{i=1}^{n}\prod_{j=1}^{i-1}\left[\mu(t_i-t_j)e^{-\int_{t_j}^{t_i}\mu(s-t_j)ds}\right]^{Z_{i,i}1_{\{I[i]=j\}}} \prod_{i=1}^{n}\prod_{j=J[i]}^{i-1}\left[\eta f(t_i-t_j)e^{-\int_{t_j}^{t_i}\eta f(s-t_j)ds}\right]^{Z_{i,j}}, \quad (2.7)$$

which is the product of marginal (inter-event time) densities $g(\cdot)$ defined by expression (1.2) for independent IPP (see also expression (1.10) in Appendix A1). The first term in square brackets is the immigrant density (1.2) and the index $I[i]$ is defined in (2.2). When a lag $t_i - t_j, j < i = 1, \ldots, n$ is an immigrant inter-event time (i.e., $Z_{i,i}1_{\{I[i]=j\}} = 1$), then $g(\cdot)$ is evaluated at that lag. In EM0,

the immigration is memoryless and thus one does not need to keep track of the previous immigrant point. The term in the second square brackets is the *offspring inter-event time density*. When a lag $t_i - t_j,\ j < i = 1, \ldots, n$ is a parent-child inter-event time (i.e., $Z_{i,j} = 1$), then the offspring inter-event time density is evaluated at that lag. To avoid undefined values of (2.7), the offspring inter-event time density is only evaluated at lags within the support of the offspring density $f(\cdot)$. This is done by defining the index function,

$$J[i] := \min(j \in \{1, \ldots, i-1\} \ : \ f(t_i - t_j) > 0), \tag{2.8}$$

which for every point t_i returns the index of the most distant previous point t_j having t_i in the support of $f(t - t_j)$. This issue is not present for $g(\cdot)$ since the immigration intensity (1.2) never vanishes.

With (2.7), the log-likelihood (2.5) can be easily derived. From here, with (1.12), we compute

$$\begin{aligned} Q_1(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{n\times n}, \hat{\boldsymbol{\theta}}^{[m]}) \ &= \ \mathrm{E}_{\boldsymbol{Z}_{n\times n}|\boldsymbol{t}_{1:n},\hat{\boldsymbol{\theta}}^{[m]}}\Big[\log \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{n\times n})\Big] \propto \\ &\quad \left[\sum_{i=1}^{n}\sum_{j=J[i]}^{i-1} \Pr[Z_{i,j} = 1|\boldsymbol{t}_{1:i}, \widehat{\boldsymbol{\theta}}^{[m]}] \log \eta f(t_i - t_j) - \int_0^r \Phi(s|\boldsymbol{t}_{1:N(s)})ds\right] \\ &\quad + \left[\sum_{i=1}^{n}\sum_{j=1}^{i-1} \Pr[Z_{i,i}1_{\{I[i]=j\}} = 1|\boldsymbol{t}_{1:i}, \widehat{\boldsymbol{\theta}}^{[m]}] \log \mu(t_i - t_j)\right. \\ &\quad \left. - \sum_{i=1}^{n+1}\sum_{j=1}^{i-1} \Pr[I[i] = j|\boldsymbol{t}_{1:i}, \widehat{\boldsymbol{\theta}}^{[m]}] \int_{t_j}^{t_i} \mu(s - t_j)ds\right]. \end{aligned} \tag{2.9}$$

For uniformity of notation, the starting and stopping times are denoted as points $t_0 = 0$ and $t_{n+1} = r$ respectively , but not included in the sample.

The E-step involves evaluating Q_1 (2.9). This requires the probabilistic definition of the branching structure. We denote the probability weights as:

$$\pi_{i,j}^{[m]} = \Pr(Z_{i,j} = 1|\boldsymbol{t}_{1:i}, \widehat{\boldsymbol{\theta}}^{[m]}), \tag{2.10}$$

and introduce the abbreviation $\pi_i^{[m]} = \pi_{i,i}^{[m]}$, for immigrants. By definition, the weights sum to one: $\sum_{j=1}^{i} \pi_{i,j}^{[m]} = 1$, $i = 1, .., n$. The first event ($i = 1$) has $\pi_1^{[m]} = \pi_{1,1}^{[m]} = 1$ and is thus an immigrant. The second event ($i = 2$) has $\pi_{2,2} + \pi_{2,1} = 1$ and thus can either be an immigrant or an offspring with the respective probabilities (2.10). Each next event has one more parameter in the probability distribution

than its predecessor. The probabilities are presented as a lower-triangular matrix $\mathbf{\Pi}_{n\times n}$ that is, at each iteration of the EM algorithm, equal to the expected value of the branching structure matrix:

$$\mathbf{\Pi}^{[m]}_{n\times n} = \mathrm{E}[\boldsymbol{Z}_{n\times n}|\boldsymbol{t}_{1:n}, \hat{\boldsymbol{\theta}}^{[m]}]. \tag{2.11}$$

Finally, we denote conditional probability weights: $\pi^{[m]}_{i,j|k} = \Pr[Z_{i,j} = 1|\boldsymbol{t}_{1:i}, I[i] = k, \hat{\boldsymbol{\theta}}^{[m]}]$, $j \leq i$ that are abbreviated $\pi^{[m]}_{i|k} = \pi^{[m]}_{i,i|k}$ for immigrants. In this notation, probabilities in (2.9) can be written as:

$$\begin{aligned}
&\Pr[Z_{i,j} = 1|\boldsymbol{t}_{1:i}, \hat{\boldsymbol{\theta}}^{[m]}] = \pi^{[m]}_{i,j},\\
&\Pr[I[i] = j|\boldsymbol{t}_{1:i}, \hat{\boldsymbol{\theta}}^{[m]}] := \omega^{[m]}_{i,j} = \pi^{[m]}_{j}\bar{\pi}^{[m]}_{j+1|j}\ldots\bar{\pi}^{[m]}_{i-1|j}, \text{ and}\\
&\Pr[Z_{i,i}1_{\{I[i]=j\}} = 1|\boldsymbol{t}_{1:i}, \hat{\boldsymbol{\theta}}^{[m]}] = \Pr[I[i] = j|\boldsymbol{t}_{1:i}, \hat{\boldsymbol{\theta}}^{[m]}]\pi^{[m]}_{i|j} = \omega^{[m]}_{i,j}\pi^{[m]}_{i|j},
\end{aligned} \tag{2.12}$$

where we have introduced weights $\omega^{[m]}_{i,k}$ and the bar denotes the complementary probability: $\bar{\pi}^{[m]}_{i,j|k} = 1 - \pi^{[m]}_{i,j|k}$. The first line of (2.12) is just the definition (2.10). The second line defines the probability that j is the last immigrant in the series of i events up to time t_i as the product of the probability π_j that j is an immigrant times the probabilities that all following events are not immigrants, all conditional on j being an immigrant. The third line defines the probability that j is the last immigrant before immigrant i in the series of i events up to time t_i. EM0 for Hawkes processes does not require any conditional probability weights. According to the thinning property [62], the probability that an observed event t_i comes from one of the independent (sub-)processes is equal to that process' share of the total CIF at t_i. Thus, conditional probability weights $\pi^{[m]}_{i,j|k}$ can be calculated using the *complete data CIF* (2.1), whereas unconditional probability weights $\pi^{[m]}_{i,j}$ require the so-called *incomplete data CIF*,

$$\lambda_*(t_i|\boldsymbol{t}_{1:N(t)}, \boldsymbol{\theta}) = \mu_*(t_i|\boldsymbol{t}_{1:N(t)}, \boldsymbol{\theta}) + \Phi(t_i|\boldsymbol{t}_{1:N(t)}, \boldsymbol{\theta}), \tag{2.13}$$

where the *incomplete data CIF of immigration* is given by a mixture of immigrant intensities:

$$\mu_*(t|\boldsymbol{t}_{1:N(t)}, \boldsymbol{\theta}) = \sum_{j=1}^{N(t)} \omega_{N(t),j} \cdot \mu(t - t_j|\boldsymbol{\theta}), \tag{2.14}$$

with weights $\omega_{N(t),j}$ (2.12) being equal to the probability that t_j is the most recent immigrant before $t_{N(t)}$. Finally, the estimation of probability weights for given observed data $\boldsymbol{t}_{1:n}$ and parameters $\hat{\boldsymbol{\theta}}$ can

be written:

$$\pi_i^{[m]} = \frac{\mu_*(t_i|\boldsymbol{t}_{1:N(t)}, \hat{\boldsymbol{\theta}}^{[m]})}{\mu_*(t_i|\boldsymbol{t}_{1:N(t_i)}, \hat{\boldsymbol{\theta}}^{[m]}) + \Phi(t_i|\boldsymbol{t}_{1:N(t)}, \hat{\boldsymbol{\theta}}^{[m]})}, \qquad \pi_{i|k}^{[m]} = \frac{\mu(t_i - t_k|\hat{\boldsymbol{\theta}}^{[m]})}{\mu(t_i - t_k|\hat{\boldsymbol{\theta}}^{[m]}) + \Phi(t_i|\boldsymbol{t}_{1:N(t_i)}, \hat{\boldsymbol{\theta}}^{[m]})}, \quad k < i = 2:N$$

$$\pi_{i,j}^{[m]} = \frac{\hat{\eta}^{[m]}\hat{h}^{[m]}(t_i - t_j)}{\mu_*(t_i|\boldsymbol{t}_{1:N(t_i)}, \hat{\boldsymbol{\theta}}^{[m]}) + \Phi(t_i|\boldsymbol{t}_{1:N(t_i)}, \hat{\boldsymbol{\theta}}^{[m]})}, \qquad \pi_{i,j|k}^{[m]} = \frac{\hat{\eta}^{[m]}\hat{h}^{[m]}(t_i - t_j)}{\mu(t_i - t_k|\hat{\boldsymbol{\theta}}^{[m]}) + \Phi(t_i|\boldsymbol{t}_{1:N(t_i)}, \hat{\boldsymbol{\theta}}^{[m]})}, \quad j, k < i = 2:N. \tag{2.15}$$

Probability weights $\pi_{i,j}^{[m]}$ and $\omega_{i,j}^{[m]}$ can be jointly computed in the following recursive way. For each event t_i, we denote the probability weight vectors $\boldsymbol{\pi}_i = (\pi_{i,1}, \ldots, \pi_{i,i})$ and $\boldsymbol{\omega}_i = (\omega_{i,1}, ..., \omega_{i,i-1})$. The first event is set to be an immigrant ($\pi_{1,1} = 1,\ \omega_{2,1} = 1$). Then the second ($\boldsymbol{\omega}_2$) and further weight vectors are computed with the following relation:

$$\begin{aligned} \boldsymbol{\omega}_i &= \left(\pi_1\bar{\pi}_{2|1}\ldots\bar{\pi}_{i-1|1}\ ,\ \ldots\ ,\ \pi_j\bar{\pi}_{j+1|j}\ldots\bar{\pi}_{i-1|j}\ ,\ \ldots\ ,\ \pi_{i-1}\right) \\ &= \left(\left(\boldsymbol{\omega}_{i-1} \circ (\bar{\pi}_{i-1|1}, ..., \bar{\pi}_{i-1|j}, ..., \bar{\pi}_{i-1|i-2})\right),\ \pi_{i-1}\right). \end{aligned} \tag{2.16}$$

This recursive equation (2.16) expresses that the weight vector $\boldsymbol{\omega}_i$ is the Hadamard product (e.g., $(a, b) \circ (c, d) = (ac, bd)$) of the previous weight vector $\boldsymbol{\omega}_{i-1}$ and a vector of complement probabilities; and with π_{i-1} concatenated to the end of the product. Thus, to compute $\boldsymbol{\omega}_i$, one uses weight vector $\boldsymbol{\omega}_{i-1}$ and computes the necessary probability weights π via (2.15). Repeating this procedure for $i = 2, \ldots, n$, produces the needed probability weights $\pi_{i,j}^{[m]},\ \omega_{i,j}^{[m]},\ i = 1, \ldots, n,\ j = 1, ..., i$.

Now we consider the M-step. Given the probability weights from the E-step, Q_1 (2.9) is maximized to obtain new estimates $\hat{\boldsymbol{\theta}}^{[m+1]}$. Because Q_1 is decomposed into immigration ($\mu(\cdot)$) and offspring ($\eta f(\cdot)$) parts without common parameters, these parts are estimated independently. The offspring part is the same as in EM0, but the immigrant part is more complicated. Estimation of parameters $\hat{\boldsymbol{\theta}}_g^{[m+1]}$ of $\mu(\cdot)$ requires maximization of the part in the second square brackets in Q_1, which is nothing more than a MLE for the immigrant inter-event time density (1.2):

$$\hat{\boldsymbol{\theta}}_g^{[m+1]} = \arg\max_{\boldsymbol{\theta}_g} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \omega_{i,j}^{[m]} \pi_{i|j}^{[m]} \log \mathrm{g}(t_i - t_j; \boldsymbol{\theta}_g), \tag{2.17}$$

with sample weights $\omega_{i,j}^{[m]}\pi_{i|j}^{[m]}$ denoting the probability that t_j and t_i are immigrants, with no other immigrant events between them (2.12). A non-parametric estimate is possible, but numerical stability

issues arise when computing $\mu(\cdot)$, as the denominator in (1.2) becomes very small. The explicit MLE for the branching ratio parameter η can be obtained by analytically maximizing (2.9) with respect to η:

$$\widehat{\eta}^{[m+1]} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\pi_{i,j}^{[m]}}{\sum_{i=1}^{n}\widehat{F}^{[m]}(r-t_i)} = \frac{n-\sum_{i=1}^{n}\pi_{i}^{[m]}}{\sum_{i=1}^{n}\widehat{F}^{[m]}(r-t_i)}, \tag{2.18}$$

where $\widehat{F}^{[m]}$ denotes the estimation of the offspring CDF: $\widehat{F}^{[m]}(t) := \int_0^t f(s|\hat{\boldsymbol{\theta}}_f^{[m]})ds$. The estimated branching ratio (2.18) is the ratio of the expected number of offspring to a number that is less than or equal to the total number n. Thus, in a finite window, the denominator inflates the estimate to account for unobserved offspring expected to occur after the end time r. Estimation of the offspring density $f(t;\boldsymbol{\theta}_f)$ is given by the density MLE with an i.i.d weighted sample:

$$\hat{\boldsymbol{\theta}}_f^{[m+1]} = \arg\max_{\boldsymbol{\theta}_f}\sum_{i=1}^{n}\sum_{j=J[i]}^{i-1}\pi_{i,j}^{[m]}\log \mathrm{f}(t_i-t_j;\boldsymbol{\theta}_f^{[m]}), \tag{2.19}$$

where the sample weights $\pi_{i,j}^{[m]}$ are the probability that t_j is parent to t_i (2.15). Here non-parametric estimation of the offspring density $f(\cdot)$ is straightforward.

2.3.2 The Semi-Complete-Data EM Algorithm (EM2)

In this section, we propose a new alternative EM algorithm for RHawkes (EM2), which is based on a reduced set of missing data. This modification significantly improves computational efficiency and memory requirements, which allows EM estimation of the RHawkes process on large datasets. The same simplification can also be applied to the classical estimation of the standard Hawkes model (EM0), and provides a great increase in computational efficiency.

Within the EM2 algorithm, we reduce the missing data to only the diagonal elements of the branching matrix $\{Z_{i,i}\}_{i=1,\dots,n}$; i.e. only indicating which points are immigrants, and which are offspring, and not indicating the parents of the offspring. This diagonal is abbreviated by $\boldsymbol{Z}_{1:n}$ and called the *immigrant vector*. Following a similar derivation to that of Q_1 (2.9) (see Appendix A1 for details), the Q

function with the *semi-complete data* $\{\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{1:n}\}$ can be written

$$\begin{aligned} Q_2(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{1:n}, \hat{\boldsymbol{\theta}}^{[m]}) &= \mathrm{E}_{\boldsymbol{Z}_{1:n}|\boldsymbol{t}_{1:n}, \hat{\boldsymbol{\theta}}^{[m]}}\Big[\log \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{1:n})\Big] \propto \\ &\sum_{i=1}^{n}(1-\pi_i^{[m]}) \log \Phi(t_i|\boldsymbol{t}_{1:N(t_i)}) - \int_0^r \Phi(s|\boldsymbol{t}_{1:N(s)})ds \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{i-1} \pi_i^{[m]}\omega_{i,j}^{[m]} \log \mu(t_i - t_j) - \sum_{i=1}^{n+1}\sum_{j=1}^{i-1} \omega_{i,j}^{[m]} \int_{t_j}^{t_i} \mu(s-t_j)ds, \end{aligned} \tag{2.20}$$

where $P[Z_i = 1] = 1$ if $\Phi(t_i|t_{1:N(t_{[}i])}) = 0$. For uniformity of notation, we denote $t_0 = 0$ and $t_{n+1} = r$ as the starting and ending times respectively where both of these points are excluded from the sample. In (2.20), similarly to EM1, the immigration and offspring processes are separated, and the immigration part is the same as in (2.9). But, unlike in EM1 (and EM0), the individual offspring processes are not decoupled. For the E-step, the weights are computed using (2.15). For RHawkes, the probabilities $\pi_i^{[m]}$ and $\pi_{i|k}^{[m]}$ $(i = 1, \ldots, n;\, k = 1, \ldots, i-1)$ are needed, but $\pi_{i,j}^{[m]}$ $(i = 1, \ldots, n;\, j = 1, \ldots, i)$ are not. For the (standard) Hawkes model, only $\pi_i^{[m]}$ $(i = 1, \ldots, n)$ are needed. In the M-step, estimation of $\mu(\cdot)$ and η is identical to EM1 and given by (2.17) and (2.18). But in contrast to EM1, the estimation of $f(\cdot)$ should be done by numerically maximizing (2.20) with respect to the parameters of $f(\cdot)$. While this method is more useful for parametric estimation (see Section 2.5.4), the non-parametric estimation of $f(\cdot)$ is more difficult than for EM1, since unit mass and positivity must be enforced.

2.3.3 Convergence of Hawkes EM Algorithms

The EM algorithm produces an *EM sequence* of parameter estimates $\{\hat{\boldsymbol{\theta}}^{[m]}\}$, belonging to the subspace $\Theta^{[0]}$ of the parameter space Θ:

$$\Theta^{[0]} = \{\boldsymbol{\theta} \in \Theta : L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}^{[0]})\}, \tag{2.21}$$

defined for an initial estimate $\hat{\boldsymbol{\theta}}^{[0]}$. The main convergence result of the EM algorithm [295] is that, provided mild regularity conditions, the EM sequence converges to a *stationary value* $\boldsymbol{\theta}^*$, which means that for the EM sequence the incomplete-data likelihood converges monotonously to an extrema (local maxima, global maxima, or saddlepoint) at $\boldsymbol{\theta}^*$. This result does not hold when the sequence of estimates reaches the boundary of the parameter space Θ.

Reasonable specifications of the Hawkes model, with or without renewal immigration, satisfy the

regularity conditions of [295]. Further, by construction, the EM sequence produced by EM0, EM1, and EM2 satisfy the probabilistic constraints of the Hawkes model ($0 < \widehat{\eta}^{[m+1]} < 1$ in (2.18)) and thus estimates do not diverge to the boundary. Thus, for convergence, it is sufficient that an initial estimate $\boldsymbol{\theta}^{[0]}$ is not at the boundary. Indeed, if one starts with the estimate $\eta^{[0]} = 0$, then from (2.12) it is clear that $\forall m > 0,\ \eta^{[0]} = 0$. Starting at $\eta^{[0]} > 0$, then obtaining the value $\eta^{[m]} = 0,\ m > 0$ in the EM sequence is only possible if the support of the offspring density $f^{[m]}$ is smaller than the smallest inter event time in the dataset (2.15),(2.18). With MLE, this could only happen if the initial estimate $f^{[m]}$ was chosen with an upper endpoint less than the smallest inter-event time. On the other hand, starting from $\eta^{[0]} = 1$ (the stationary upper boundary) and $\mu^{[0]} > 0$, it is clear from (2.12)–(2.18) that $\mu^{[1]} > 0$ and $\eta^{[1]} < 1$, and by induction that $\mu^{[m]} > 0$ and $\eta^{[m]} < 1$, $\forall m > 0$. Thus, convergence issues should be avoided by taking $\eta^{(0)} > 0$ and large enough upper endpoint of $f^{(0)}$ such that $\inf(t \in \mathbb{R} : F^{[0]}(t) = 1) > \min(\{t_i - t_{i-1}, i = 1, ..., n\})$. However, since neither the Hawkes nor RHawkes likelihoods are necessarily unimodal (see [93] for an example), it cannot be guaranteed that the stationary value $\boldsymbol{\theta}^*$ is at the global maximum. Thus it is recommended to take multiple starting estimates $\boldsymbol{\theta}^{[0]}$ and select the best estimate from the resulting sequences.

Now we consider the speed of convergence. The EM Algorithm implicitly defines a mapping $M : \boldsymbol{\theta}^{[m]} \to \boldsymbol{\theta}^{[m+1]}$. In [71], it was shown that (i) all EM algorithms have (at least) linear convergence with rate $dM/d\boldsymbol{\theta}$, and (ii) that the rate of convergence near the local optimum $\boldsymbol{\theta}^*$ is fast when the proportion of the Fisher information in the missing data to that of the complete data is small. For this reason, EM2, with reduced missing data, may have faster convergence than EM1. However, for RHawkes, this cannot be shown explicitly due to the multiplicative form of immigrant probabilities in (2.17) and thus the complexity of the closed form of $dM/d\boldsymbol{\theta}$. Note that an example of the case with large missing information is when clusters are overlapping and thus the branching structure is highly uncertain. Further, often it is possible to formulate the EM algorithm as a projected gradient ascent method [218]:

$$\boldsymbol{\theta}^{[m+1]} - \boldsymbol{\theta}^{[m]} = P(\boldsymbol{\theta}^{[m]}) \nabla \mathrm{logL}|_{\boldsymbol{\theta}^{[m]}}, \tag{2.22}$$

where one takes a positive step in the direction of the projected gradient of the log-likelihood. As an example, in [164], the EM0 for a Hawkes model was presented in this form. The projected gradient ascent method (2.22) is a first order algorithm on a locally reshaped log-likelihood surface. In [218], it was shown that, when the missing information is small compared to the complete information, EM

exhibits approximate Newton behaviour (superlinear convergence) near $\boldsymbol{\theta}^*$. It was claimed to hold for all latent variable models (this includes RHawkes).

In agreement with [297, 218, 260], simulations in [164] confirmed that, for the Hawkes process with well separated clusters, the EM algorithm behaved like a second order method, whereas for overlapping clusters the convergence was closer to linear. These results are encouraging for the EM algorithms presented here. However the EM algorithms here have not been verified to fit into the form of (2.22) due to the difficulty of expressing the partial derivatives of the log likelihood given by a logarithm of a sum of products (see expression (2.4) in Appendix A1). Heuristically speaking, over-dispersed immigration in RHawkes can worsen cluster overlap and thus decrease the speed of convergence. For the sake of brevity, we do not present analysis of the convergence of EM1 and EM2 for RHawkes process.

A practical convergence issue was also encountered in some circumstances. Specifically, when estimating RHawkes with EM2 on data with branching ratio ≈ 0, the scale of the offspring density may be overestimated with each successive iteration, leading to an erroneously large branching ratio estimate due to the denominator of (2.18). Thus, to avoid this error, one may wish to set the denominator of (2.18) to the sample size n, accepting a negative bias for the estimation of η, which vanishes with increasing sample size, and is negligible for an offspring density with short memory. This issue was not encountered when estimating the Hawkes process with IPP immigration with EM2.

2.3.4 Computational Efficiency for Estimation of the Hawkes Process

Here the computational complexity and memory requirements for the estimation of Hawkes and RHawkes models are discussed. First let us note that in general both MLE and the E-step for all EM algorithms, require evaluating the CIF at $\mathcal{O}(n^2)$ inter-event times. However in the case of parametric estimation with an exponential offspring density (or any linear mixture of exponentials), a recursive relationship [197] can be used for reducing the computational complexity to $\mathcal{O}(n)$ for MLE and EM2 for Hawkes, including Hawkes with IPP immigration.

The memory requirements for storing the probability weights in the E-step are the following: $\mathcal{O}(n^2)$ in EM0, EM1 and EM2 for RHawkes; and $\mathcal{O}(n)$ in case of EM2 for Hawkes. For the M-step, weighted samples of $O(n^2)$ inter-event lags are used for the offspring density (2.19), and also for the renewal density (2.17) in the RHawkes case. For EM0 and EM1, this poses only a minor problem, because if

these samples are too large, then subsamples can be taken. Such a sub-sampling approach also allows for the use of density estimators that do not naturally include sample weights. This approach is easier than the variational calculus approach used in [165].

The computational difficulty can be further reduced if the offspring density has finite support with upper endpoint t_f and $n_f = \max(\{N(t_i) - N(t_j) : t_i - t_j < t_f)\})$ being the largest number of points observed within the support of the density. In this case, the E-step and M-step only need to be performed on lags $\{t_i - t_j : i = 1, \ldots, n, j = \max(1, i - n_f), \ldots, i - 1\}$. The same is true for the renewal density. This reduces both $O(n^2)$ memory and computation requirements to $O(n \cdot n_f)$ with $n_f \leq n$. A similar approach, which adaptively chooses n_f to satisfy a pre-specified tolerable error, was introduced in [111]. Note that taking n_f too small introduces downward bias into the estimation of the branching ratio (2.18).

2.4 Statistical Inference

Here we discuss parameter variance-covariance, likelihood, and p-value computation for the RHawkes process. Without a closed-form solution, estimation of the parameter variance-covariance matrix should be done via bootstrap. As for similar models (such as mixture models), the sample size must be large to achieve the asymptotic covariance, and Monte Carlo estimates are typically recommended (see [103] and references therein).

Likelihood and goodness of fit tests for RHawkes require evaluating the CIF (2.1), and thus may be computed for each immigrant vector $\boldsymbol{Z}_{1:n} = \{Z_{i,i}\}_{i=1,\ldots,n} \in \{0,1\}^n$. Since the first point of the sample is always set to be an immigrant, there exist 2^{n-1} possible immigrant vectors, uniquely indexed as $\boldsymbol{z}_{1:n}^{(j)},\ j = 1, \ldots, 2^{n-1}$.

To simplify computation, a Monte Carlo approach can be used where sample averages of likelihoods and p-values are taken. Specifically, given the probabilistic description of the branching structure obtained in the E-step (2.12), an ensemble of realizations of the immigrant vector may be generated, and likelihoods and p-values computed for each realization. Ensemble average likelihoods and p-values may then be computed. A Monte Carlo study of the inferential power of these statistics is conducted in Section 2.5.

To simulate realizations $\boldsymbol{z}_{1:n} = (z_1, \ldots, z_n)$ of the random vector $\boldsymbol{Z}_{1:n} \in \{0,1\}^n$, we have used an acceptance-rejection thinning type algorithm [167]. We start from the vector $\boldsymbol{z}_{1:n} = (1, 0, \ldots, 0)$

since the first point is always treated as an immigrant. Next, for each following event t_i, $i = 2, \ldots, n$, Bernoulli random variables with probabilities $\pi_{i|1}$ (2.15) are generated, and the first success (at $i = k$) is taken as the second immigrant (i.e. $z_k = 1$); then the third immigrant is selected in the same way with probabilities $\pi_{i|k},\ i = k+1, \ldots, n$; and so on until the stopping time r is reached. This yields a realization $\boldsymbol{z}_{1:n} = \boldsymbol{z}_{1:n}^{(a)}$, identified by an index $a \in \{1, \ldots, 2^{n-1}\}$. Repeating the procedure l times, we obtain the sample set $\{a_i\}_{i=1,\ldots,l}$, where each element defines the sampled immigrant vector $\boldsymbol{z}_{1:n}^{(a_i)}$.

2.4.1 Likelihood

The likelihood value for RHawkes was computed as follows: For a given immigrant vector $\boldsymbol{z}_{1:n}^{(j)}$, $j \in \{1, \ldots, 2^{n-1}\}$, the immigration intensity,

$$\mu^{(j)}(t) = \mu(t - t_{I[N(t)]} | \boldsymbol{z}_{1:n}^{(j)}) \ , \tag{2.23}$$

is a deterministic function, and the RHawkes model can be treated as a Hawkes model with IPP immigration. Thus, plugging the Hawkes CIF (1.4) with immigration intensity (2.23) into the log-likelihood equation (1.9), one obtains the *conditional incomplete data likelihood* for immigrant vector $\boldsymbol{z}_{1:n}^{(j)}$:

$$\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{t}_{1:n} | \boldsymbol{z}_{1:n}^{(j)}) = \prod_{i=1}^{n} \left(\mu^{(j)}(t_i) + \Phi(t_i | \boldsymbol{t}_{1:N(t_i)}) \right) \exp\left(- \int_0^r \mu^{(j)}(s) + \Phi(s | \boldsymbol{t}_{1:N(s)}) ds \right). \tag{2.24}$$

The incomplete data likelihood is equal to a weighted sum of the conditional incomplete likelihoods (2.24):

$$\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{t}_{1:n}) = \sum_{j=1}^{2^{n-1}} \mathrm{L}(\boldsymbol{\theta}; \boldsymbol{t}_{1:n} | \boldsymbol{z}_{1:n}^{(j)}) \Pr[\boldsymbol{Z}_{1:n} = \boldsymbol{z}_{1:n}^{(j)} | \boldsymbol{\theta}]. \tag{2.25}$$

The weighting probabilities in (2.25) may be computed by probabilities from the E-step (2.12), however, this is computationally burdensome. Instead, a Monte Carlo approximation of the likelihood (2.25),

$$\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{t}_{1:n}) \approx \frac{1}{l} \sum_{i=1}^{l} \mathrm{L}\left(\boldsymbol{\theta}; \boldsymbol{t}_{1:n} | \boldsymbol{z}_{1:n}^{(a_i)} \right), \tag{2.26}$$

may be done with sampled immigrant vector indices $\{a_i\}_{i=1,\ldots,l}$.

Ultimately, the Monte Carlo log-likelihood is obtained by taking the logarithm of this average (2.26). The logarithm of the incomplete likelihood (2.25) or its approximation (2.26) may be directly compared

with the log-likelihood of the standard Hawkes process (1.9). In practice, one will start by calculating the log of (2.24), and exponentiate these to be averaged in (2.25). Due to limitations of floating-point number representation, this exponentiation may result in an overflow being evaluated by the computer. In this case, one can average the log of the conditional likelihoods (2.24), keeping in mind that this will provide an underestimate of the Monte Carlo log-likelihood due to Jensen's inequality. The underestimation error will be small when the variance in the conditional log-likelihoods is small with respect to the mean.

2.4.2 p-Values

To perform a hypothesis test for an estimated point process model, one often uses the so-called *residual analysis* [193] based on the *time change property* [198]: For point process $\{T_i\}_{i\in\mathbb{N}}$ with CIF $\lambda(t|\boldsymbol{t}_{1:N(t)})$, the set of *transformed times* $\{\tilde{T}_i\}_{i\in\mathbb{N}}$, $\tilde{T}_i = \int_0^{T_i} \lambda(s|\boldsymbol{t}_{1:N(s)})ds$ are generated by a unit rate Poisson process. Thus for a realization $\boldsymbol{t}_{1:n}$, one can estimate its CIF, transform it to $\tilde{\boldsymbol{t}}_{1:n}$ and test the hypothesis that the resultant process is unit Poisson. More generally, we define the *test statistic* (for example, the Kolmogorov Smirnov (KS) distance [179]) as a random variable $S := S(\boldsymbol{T}_{1:n}, \boldsymbol{Z}_{1:n})$, which, under the null hypothesis, has known *reference distribution* F_0. Here the observed test statistic $S(\boldsymbol{t}_{1:n}, \boldsymbol{z}_{1:n}^{(j)})$ transforms a realization of points $\boldsymbol{t}_{1:n}$, given their immigrant vector $\boldsymbol{z}_{1:n}^{(j)}$, $j \in \{1, \ldots, 2^{n-1}\}$. For semi-complete data sets $\{\boldsymbol{t}_{1:n}, \boldsymbol{z}_{1:n}^{(j)}\}$, we define the null hypothesis $H_0^{(j)}$ as the validity of the RHawkes model for $\boldsymbol{Z}_{1:n} = \boldsymbol{z}_{1:n}^{(j)}$. Then the *semi-complete data p-values* are given by

$$p^{(j)} = \Pr[S > S(\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{1:n})|H_0^{(j)}] = 1 - F_0(S(\boldsymbol{t}_{1:n}, \boldsymbol{z}_{1:n}^{(j)})),\ j = 1, \ldots, 2^{n-1}. \tag{2.27}$$

For the incomplete data set $\{\boldsymbol{t}_{1:n}\}$, the null hypothesis H_0 is that the RHawkes model is true. The test statistic for this, $S(\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{1:n})$, is unknown because the immigrant vector is unknown. Thus, by conditioning, the *incomplete data p-value* is,

$$p = \Pr[S > S(\boldsymbol{t}_{1:n}, \boldsymbol{Z}_{1:n})|H_0] = \sum_{j=1}^{2^{n-1}} p^{(j)} \Pr[\boldsymbol{Z}_{1:n} = \boldsymbol{z}_{1:n}^{(j)}|\boldsymbol{\theta}], \tag{2.28}$$

which may be expressed in terms of the semi-complete data p-values (2.27). As for the likelihood, a Monte-Carlo approximation of the p-value (2.28) may be done by taking the average of the semi-

complete data p-values:

$$p \approx \frac{1}{l} \sum_{i=1}^{l} p^{(a_i)}, \tag{2.29}$$

having their indices in the sampled set $\{a_i\}_{i=1,\dots,l}$.

2.5 Monte Carlo Study of the EM estimation of RHawkes

In this section, we examine the convergence (consistency) of EM1, the performance of the statistical inference methods developed in Section 2.4, and the robustness of the Hawkes process estimation in the case of misspecification of the immigration process. For the study of the EM2 algorithm, we refer to Appendix A2. We consider the renewal process immigration with a Weibull immigrant waiting time distribution:

$$g(w) = \mu(w) \exp\left(- \int_0^w \mu(s) ds \right) = \frac{\kappa}{\beta} \left(\frac{w}{\beta} \right)^{\kappa - 1} \exp\left(- \left(\frac{w}{\beta} \right)^{\kappa} \right), \tag{2.30}$$

and associated Weibull intensity function $\mu(w) = \kappa w^{\kappa-1}/\beta^{\kappa}$, $w \geq 0$. The case $\kappa = 1$ corresponds to the standard Hawkes process (1.4) with constant background intensity $\mu = 1/\beta$. When $\kappa < 1$, the intensity decays, which implies that the inter-event time density (2.30) is sub-exponential and features over-dispersion. Alternatively, as $\kappa \to \infty$, the inter-event density (2.30) weakly converges to a delta-function $g(w) = \delta(w - \beta)$, and the immigration process becomes deterministic with regular event spacing β.

For the offspring density $f(t)$, we consider both the exponential pdf, originally suggested by Hawkes [117]:

$$f_{exp}(t) = \frac{1}{\tau_0} \exp\left(-\frac{t}{\tau_0} \right) 1_{t \geq 0}, \tag{2.31}$$

which is parametrized with a shape parameter $\tau_0 > 0$; and the Omori-type heavy-tailed pdf [193]:

$$f_{Omori}(t) = \frac{\alpha c^{\alpha}}{(t + c)^{1+\alpha}} 1_{t \geq 0}, \tag{2.32}$$

with shift parameter $c > 0$ and Pareto tail index $\alpha > 0$.

The exponential offspring density (2.31), which is typical for financial and econometric applications [34, 29, 91, 80, 11], endows Markov properties to the model [190], and is more robust to outliers than heavy-tailed alternatives [93]. A heavy-tailed offspring density (2.32) is typical for seismological applications [194], where it accounts for the power law decay of aftershock activity with time (*Omori's*

law). For many practical applications, it can be well approximated as a sum of weighted exponential densities [112]. As discussed in Section 2.3.4, the computational complexity of evaluation of the log-likelihood in the above cases can be reduced to $\mathcal{O}(n)$.

2.5.1 Bias and Efficiency

This section discusses the bias and efficiency of the EM1 estimator for RHawkes (2.1). For this, we have considered simulations of the RHawkes process (2.1) with parameters κ and η presented in Table 2.1. The Weibull shape parameter κ was given values in $\{0.5, 0.75, 1, 1.25, 1.5\}$, ranging from highly over-dispersed to highly under-dispersed. For each value of κ, the scale parameter β was chosen such that the expected immigrant inter-event time was equal to 10, i.e., β was given values in $\{5, 8.4, 10, 10.7, 11.1\}$. The characteristic time τ_0 of the exponential offspring density (2.31) was chosen to be $\tau_0 = 3$.

For each combination of parameters, we simulated 50 independent RHawkes realizations, each with 500 events. Efficient simulation was performed using the algorithm of [184], which exploits the branching representation of the Hawkes process. The model parameters $\{\kappa,\ \beta,\ \eta,\ \tau_0\}$ were then estimated using EM1. We intentionally chose "bad" starting parameter estimates to demonstrate robust convergence: $\hat{\kappa}^{[0]} = 1$; the scale parameter $\hat{\beta}^{[0]}$ was chosen as the true value β multiplied by a uniform random number in $[0.25, 4]$; the branching ratio $\hat{\eta}^{[0]}$ was chosen as a uniform random number in $[0.1, 0.9]$; and the characteristic time of the offspring density τ_0 was chosen as a uniform random number in $[0.5, 10]$.

The bias and standard deviation of the estimates are presented in Table 2.1. In general, most parameters were well estimated, especially the branching ratio η. Due to the fixed sample size of 500 points, when η is larger, the expected number of immigrants $\mathrm{E}[N^{(0)}(r)]$ is smaller. Thus the bias and the variance of estimates of immigration process parameters $\hat{\kappa}$ and $\hat{\beta}$ are larger with larger η and are the worst for $\eta = 0.9$, i.e., when the $\mathrm{E}[N^{(0)}(r)] = 50$. Another factor that introduces systematic error into the results is that when η is large and thus clusters are overlapping. However, this bias decreases with increasing sample size.

Table 2.1: Results of EM1 estimation (Section 2.3) of the Hawkes Process with Weibull renewal immigration (2.1) and exponential offspring density on simulated data. For each combination of parameters, this table presents the average bias and standard deviation (in brackets) of estimates over 50 simulations.

κ	η	$\mathrm{E}[N_{(0)}(\tau)]$	$\widehat{\kappa}-\kappa$	$\widehat{\beta}-\beta$	$\widehat{\eta}-\eta$	$\widehat{\tau_0}-\tau_0$
0.5	0.1	450	0.02 (0.02)	0.55 (0.74)	0.02 (0.06)	-0.05 (2.04)
	0.5	250	0.06 (0.03)	1.70 (1.32)	0.03 (0.06)	-0.41 (0.52)
	0.9	50	0.16 (0.15)	0.89 (2.13)	-0.04 (0.05)	-0.46 (0.51)
0.75	0.1	450	0.04 (0.04)	1.01 (1.22)	0.06 (0.06)	0.31 (1.51)
	0.5	250	0.06 (0.07)	1.16 (1.52)	0.02 (0.06)	-0.25 (0.48)
	0.9	50	0.12 (0.13)	-1.00 (3.14)	-0.05 (0.05)	-0.52 (0.37)
1	0.1	450	0.02 (0.05)	0.46 (0.76)	0.03 (0.04)	1.71 (2.97)
	0.5	250	-0.02 (0.07)	-0.66(1.08)	-0.03 (0.05)	-0.08 (0.48)
	0.9	50	0.02 (0.15)	-1.58 (2.99)	-0.04 (0.05)	-0.28 (0.60)
1.25	0.1	450	-0.03 (0.08)	0.04 (0.63)	0.01 (0.04)	6.23 (6.59)
	0.5	250	-0.06 (0.11)	-0.69 (1.21)	-0.04 (0.07)	-0.05 (0.59)
	0.9	50	-0.12 (0.20)	-3.16 (2.53)	-0.07 (0.05)	-0.30 (0.49)
1.5	0.1	450	-0.06 (0.09)	-0.09 (0.6)	0.00 (0.03)	3.91 (5.35)
	0.5	250	-0.15 (0.10)	-0.79 (1.09)	-0.03 (0.06)	0.03 (0.56)
	0.9	50	-0.29 (0.26)	-3.17 (2.96)	-0.05 (0.05)	-0.36 (0.42)

2.5.2 Model Selection

In this section, we address the question of model selection when the immigration process is unknown. For this, we simulate the Hawkes process with Weibull renewal immigration (2.1) and exponential offspring density (2.31), and then test the null hypothesis (H_0) that observed events $\{\boldsymbol{t}_{1:n}\}$ are generated with the Hawkes model (1.4) versus the alternative hypothesis (H_1) that $\{\boldsymbol{t}_{1:n}\}$ are generated from a Hawkes process with Weibull renewal immigration (2.1). In both models (H_0 and H_1), the offspring density is assumed to be exponential (2.31).

We consider three statistical test: (i) comparison of the AIC values of H_0 and H_1, (ii) the Wilks likelihood ratio test with level 0.05 [293] where H_0 is nested in H_1, and (iii) the KS test for standard exponential transformed inter-event times with level 0.05, using the methodology discussed in Section 2.4. It should be noted that the option (iii) is not a test of H_0 against alternative H_1, but the Portmanteau-type test of H_0 against the alternative hypothesis $\tilde{H}_1$, that is loosely specified (i.e. "not H_0").

The parameters for the process (2.1), (2.30) were chosen as follows: The Weibull shape parameter

κ was given values in $\{0.5, 0.75, 1, 1.25, 1.5\}$ and other parameters were fixed at values of $\beta = 1$, $\eta = 0.6$ and $\tau_0 = 0.3$. For each combinations of these parameters, we simulated 100 independent realizations of size 250, 500, 750, and 1000 events. Then both models were estimated on each sample: the true model (2.1), which corresponds to H_1, was estimated using EM1, and the misspecified model (1.4), which corresponds to H_0, was estimated using direct maximization of the log-likelihood (1.9). The Monte-Carlo approximation of the likelihood for the true model (2.26) was done with 200 sampled likelihoods.

Table 2.2 summarizes the results. In general, the larger the sample and the further from Poisson immigration (when κ is away from 1), the more powerful the test. The AIC test (i) provides a powerful decision rule for comparing the models, even for small sample sizes (e.g., $n = 250$) and moderately over and under dispersed immigration (e.g., $\kappa = 0.75$ and $\kappa = 1.25$ respectively). When the null model is true (i.e., $\kappa = 1$), both models provide approximately equal AIC. The Wilks test (ii) is powerful for sample sizes with 500 or more points, and at even smaller sample sizes in case of high over- or under-dispersion of the immigration process (e.g., $\kappa = 0.5$ and $\kappa = 1.5$ respectively). The KS test (iii) is understandably the least powerful as it specifies no alternative model. Even on large sample sizes ($n = 1000$, $\mathrm{E}[N^{(0)}(r)] = 400$), and for significant immigrant over-dispersion. ($\kappa = 0.5$), the test has very low power: less than 0.5.

Summarizing, model selection can be successfully resolved using AIC and/or the Wilks test. In the following section, we will see that misspecification of the immigration process can bias parameter estimates.

2.5.3 Robustness of Branching Ratio Estimation under Misspecification of the Immigration Process

The branching ratio η is an important parameter of the Hawkes model and branching-type processes in general, as it quantitatively defines both the stationarity of the system and the importance of the self-exciting mechanisms. In this section, we discuss the robustness of the estimation of η when the immigration process is misspecified. For this, we estimate the Hawkes model on synthetic data generated with the RHawkes model as well as the Hawkes model with IPP immigration. Both of these examples illustrate ways in which branching ratio estimation can be systematically biased (see [93] for a longer list and discussions).

First we consider the generating process being RHawkes (2.1) with Weibull immigration (2.30) and exponential offspring density (2.31). We fixed parameters $\eta = 0.5$ and $\tau = 0.1$ and varied the immigration shape parameter from over-dispersed ($\kappa = 0.4$) to under-dispersed ($\kappa = 1.4$). As before, the scale parameter β was chosen so that, for any given κ, the expected immigrant inter-event time was fixed (in this case $\mathrm{E}[T_i^{(0)} - T_{i-1}^{(0)}] = 4$).

For each value of κ, we have simulated 50 independent realizations. On each synthetic realization, we have used MLE to estimate the (standard) Hawkes process (1.4) with (i) exponential offspring density (2.31) and (ii) Omori-type density (2.32). Figure 2.1 presents results of the estimation of the branching ratio $\hat{\eta}$ as a function of the shape parameter κ of the underlying immigration process.

In Figure 2.1 (left), both models with Poisson immigration have a significant bias in the estimation of $\hat{\eta}$. In the case of under-dispersed immigration ($\kappa > 1$), one observes a relatively small negative bias, which is similar for exponential (2.31) and Omori-type (2.32) offspring densities. In contrast, for over-dispersed immigration ($\kappa < 1$), the bias is positive and much stronger. For instance, when $\kappa = 0.5$, the branching ratio has median positive bias of 0.17 and 0.31 for the Hawkes process with exponential and Omori-type offspring densities respectively.

Next, we consider the simulation study from Appendix A2, namely EM2 estimation of the Hawkes

Test	n	$E[N^{(0)}(r)]$	$\kappa = 0.5$	$\kappa = 0.75$	$\kappa = 1$	$\kappa = 1.25$	$\kappa = 1.5$
AIC	250	100	0.99	0.58	0.06	0.35	0.81
	500	200	1	0.79	0.07	0.6	0.95
	750	300	1	0.93	0.12	0.68	0.99
	1000	400	1	0.96	0.2	0.83	1
Wilks	250	100	0.97	0.35	0.01	0.19	0.54
	500	200	1	0.67	0.01	0.32	0.87
	750	300	1	0.85	0.04	0.5	0.95
	1000	400	1	0.9	0.05	0.65	1
KS	250	100	0.08	0.05	0.03	0.06	0.1
	500	200	0.17	0.04	0.03	0.08	0.17
	750	300	0.30	0.04	0.06	0.13	0.22
	1000	400	0.46	0.06	0.05	0.14	0.22

Table 2.2: Results of model selection tests. $\mathrm{E}[N^{(0)}(r)]$ denotes the expected number of immigrant events in the sample. AIC provides the fraction of the 50 repetitions in which the H_1 model had superior AIC to the H_0 model. Wilks provides the fraction of the 50 repetitions in which the H_0 model was rejected when compared to the H_1 model using the Wilks test at level 0.05. KS provides the fraction of the 50 repetitions in which the H_0 model was rejected when using the KS test at level 0.05.

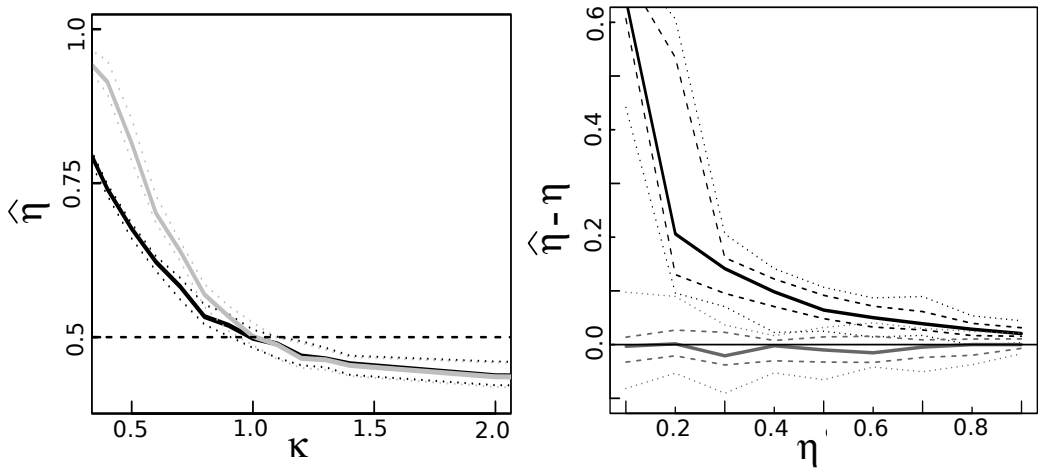

Figure 2.1: The left panel: EM1 estimates of the branching ratio $\hat{\eta}$ (2.18) using the Hawkes model with Poisson immigration and exponential offspring density (black) and Omori-type density (grey) when the true process is generated with the Hawkes model with Weibull renewal immigration with shape parameter κ. The true branching ratio (0.5) is presented with a horizontal dashed line. Solid lines correspond to median values and dotted lines present quartiles of estimates for the densities.
The right panel: The difference between the estimated branching ratio and the true branching ratio: 1) (black) where the Hawkes model is estimated on simulations from the Hawkes model with IPP immigration for a range of branching ratios, and 2) (grey) where the model used for simulation is estimated on its own realizations. The horizontal axis is the value of the branching ratio used for simulation. The median (heavy solid), quartiles (dashed), and 0.05 and 0.95 quantiles (dotted) of all estimates are given.

process with deterministic sinusoidal immigration intensity. Figure 2.1 (right) summarizes the errors in the estimation of the branching ratio for the true model (IPP immigration) and the false model (constant immigration). For the true model, estimation is consistent and efficient. For the false model, the branching ratio is consistently overestimated, in particular for low values of the branching ratio. For example, an upward bias of more than 0.6 is observed when the true branching ratio $\eta = 0.1$, and still an upward bias of approximately 0.1 when $\eta = 0.5$. The overestimation is a result of an apparent clustering provided by the baseline sinusoidal IPP immigration $\mu(t)$, which is attributed by the (standard) Hawkes model to the self-exciting term $\Phi(\cdot)$.

2.5.4 Study of EM2 Estimation of Hawkes with Inhomogeneous Poisson Immigration

Here we perform a Monte Carlo study using EM2 to estimate the (standard) Hawkes process (1.4) with deterministic IPP immigration intensity $0 < \mu(t) < \infty,\ \forall t$. For computational efficiency, an exponential offspring density (2.31) is chosen. In this case, the recurrence relation of [197] can be used and both the E- and M-steps of the EM2 have complexity of $\mathcal{O}(n)$. Thus, it becomes possible to estimate the model even on large datasets with a standard personal computer. The immigration intensity will be estimated using kernel density estimation

$$\widehat{\mu(t)} = \sum_{i=1}^{n} \pi_i k(t - t_i; b) 1_{\{0<t<r\}} + c(t), \tag{2.33}$$

where the kernel function $k(t; b)$ is a pdf with bandwidth parameter b. This estimator (2.33) distributes mass π_i around each point t_i, and the higher the bandwidth, the more dispersed the mass is. When mass is distributed outside of the interval $(0, r]$, it should be "reflected" back in. We denote this "reflection" operation by the term $c(t) \geq 0,\ 0 < t \leq r$. One of the crucial aspects of using kernel estimators like (2.33) is the selection of bandwidth. Here we use a fixed bandwidth that is specified a-priori, however for practical applications automatic bandwidth selection procedures may be used [232, 274]. When a measure of the degrees of freedom of the estimate is available, AIC tests may be used. An essential feature of this estimator (2.33) is that $\int_0^r \widehat{\mu(t)} = \sum_{i=1}^n \pi_i$, which means that it provides an unbiased estimation for the total number of immigrant points in the sample. This avoids the accumulation of bias across EM iterations.

In our Monte Carlo study we have simulated the Hawkes process with sinusoidal immigration intensity $\mu(t) = \sin(2\pi t/250) + 1.5$, exponential offspring density (2.31) with scale parameter $\tau_0 = 0.1$, and branching ratio values η sweeping from 0.1 to 0.9 by 0.1. For each set of parameters, 50 simulations of this process on one period of the immigration intensity $(0, 250]$ were performed. The median sample size was 1200 with quartiles 520 and 1310. For each realization, we estimated two models using the EM2 algorithm (see section 2.3.2): the first being the true model, and the second (false model) being the Hawkes model with homogeneous immigration ($\mu(t) = \mu$). The initial parameter estimates were chosen uniformly at random in the following intervals $\eta \in (0.1, 0.9)$, $\tau_0 \in (0.1, 10)$ and $\mu \in (0.1, 5)$. The EM algorithm was allowed to perform 200 iterations, but in 90 percent of the time it converged

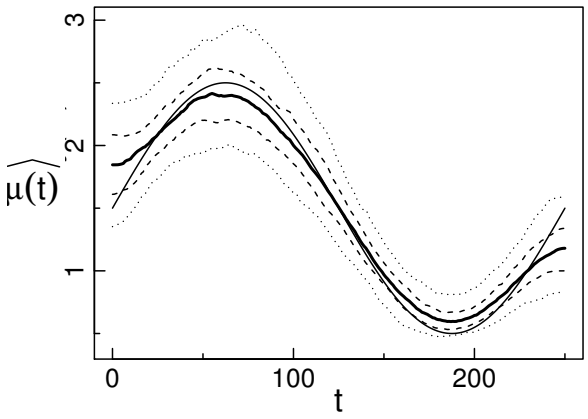

Figure 2.2: The solid thin black line is the true sinusoidal immigration intensity used in simulation. Lines are also plotted for the median (heavy solid), quartiles (dashed), and 0.05 and 0.95 quantiles (dotted) of all estimates for all parameter values.

in less than 100 iterations. The convergence criterion was that the cumulative sum of the absolute differences of estimated parameters for the previous 3 iterations were smaller than or equal to 10^{-6}.

The comparison of the estimated branching ratios $\hat{\eta}$ for true and false models was discussed in Section 2.5.3. Here we illustrate the good convergence of the non-parametric estimation of the immigration intensity $\mu(t)$. Figure 2.2 presents the true immigration intensity and a summary of the estimated immigration intensity across all samples for the true model. As it is seen from the Figure, the immigration intensity is well estimated, including cases when most of the points are offspring (when $\eta = 0.9$).

2.6 Case Study: Self-Excitation of Mid-Price Changes of the E-mini S&P500 Futures

Over the past decade, the Hawkes model has become very popular in financial applications, mostly for modeling high-frequency fluctuations of prices and limit order book dynamics [34, 29, 273, 55, 11, 91, 21, 48, 112], and for the modeling of sequences of lower-frequency extreme events [47, 84, 80]. In this section, we present a data-driven motivating example for applications of the RHawkes model within the context of quantitative finance.

We study the point process of *mid-quote price* changes of the E-mini S&P500 Futures Contract traded on the Chicago Mercantile Exchange (CME). The mid-quote price is defined as the mean of the

best bid and best ask prices, and is the standard point process employed in the study of high-frequency (sub-second) price fluctuations [33]. Here the branching ratio η is defined as the proportion of price changes that are caused by previous price changes, thus quantifying market (in-)stability concepts such as *endogeneity*, *positive-feedback* and *reflexivity*, and its estimation gives deep insight into the dynamics of the system [91].

We consider price dynamics for 10 separate trading days within 2012 (2012-01-03, 2012-02-03, 2012-03-01, 2012-04-03, 2012-05-03, 2012-06-01, 2012-08-03, 2012-09-04, 2012-10-03, 2012-11-01). Similarly to [91], we define a sample as a sequence of mid-price changes occurring in 20 minute windows, where each window overlaps 10 minutes with the previous window. Only windows occurring during the Regular Trading Hours (9:30 to 16:15 CDT) were considered. This yields almost 400 samples. Samples with less than 200 points were excluded (approximately 20 percent of samples), resulting in a median sample size of around 400 points. As discussed in [93], due to data quality issues, the points are uniformly randomized within intervals of 0.2 seconds prior to their given time-stamp.

We have estimated three models on each dataset: (M1) the Hawkes model with exponential offspring via direct MLE, (M2) the Hawkes model with IPP immigration and exponential offspring using EM2 with immigration estimated by kernel estimation (see eq. (2.33) in Appendix A2), and (M3) the RHawkes model with Weibull immigration and exponential offspring using EM1. The main result of the analysis below is that the log-likelihood of the model with renewal immigration (M3) is significantly larger than that of the models with Poisson immigration processes (both M1 and M2)

Specifically, we have compared differences between log-likelihoods of models. The (0.1, 0.25, 0.5, 0.75, 0.9) quantiles of differences between (M3) and (M1) were (5.1, 9.1, 16.3, 24.8, 34.7) and clearly favour the (M3) model. Even for the 0.1 quantile log-likelihood difference (i.e., 5.1), the Wilks likelihood ratio test rejects the simpler model (M1) with a p-value of 0.001. The AIC test also favours (M3) as the difference in numbers of parameters between the models is equal to 1.

The quantiles of differences in log-likelihoods between models (M3) and (M2) were (0.6, 4.5, 10.4, 18.6, 26.0). Since the models are not nested, Wilks-type of tests can not be performed. Further, comparison of penalized likelihoods, such as AIC, cannot be done since the *equivalent degrees of freedom* (*edf*) of the kernel estimate of (M2) are not known. However, even in the simplest case when $\mu(t)$ is constant and $edf = 1$, (M3) is superior to (M2) nearly 90 percent of the time according to AIC test.

Using the standard goodness-of-fit test based on the transformed inter-event times that should

follow exponential distribution (Section 2.4), (M1) is rejected on 53 percent of samples at a level of 0.05, (M2) on 51 percent of samples, and (M3) on 22 percent of the samples, which again indicates that (M3) is superior. However we need to note that for all models the rejection rate is quite high.

Alternatively, one may consider testing if the transformed time follows a uniform distribution on the transformed interval $(0, \tilde{r}]$ since this is another definition of a homogeneous Poisson process. Unlike the previous test, this takes into account temporal correlations between inter-event times. Here, the rejection rates at level 0.05 are equal to: (M1) 0.25, (M2) 0.08, and (M3) 0.20. Thus, (M2) does a better job of "detrending", while (M3) better describes the inter-event distribution and obtains superior likelihood. One could, for instance, first detrend the data and then apply (M3).

Finally, we summarize the parameter estimates of the winning model (M3). The branching ratio η had median 0.65 and quartiles (0.61, 0.7). These high values of η suggests substantial high-frequency self-excitation in price changes: about 65% of price changes are triggered by previous price changes. This supports, while correcting slightly downward, the previous estimation of [91] based on (M1), which provides higher estimates with a median of 0.73.

The immigration shape parameter κ had median 0.55 and quartiles of (0.50, 0.61), which consistently indicates over-dispersion. in the immigration process, relative to the Poisson. The immigration scale parameter β has median 5.2 and quartiles of (3.4, 8.2). The large variance of β reflects substantial changes of activity levels within each day. Finally, the scale parameter of the offspring density has median 0.056 and quartiles of (0.052, 0.061), that are similar for the three models.

The superiority of RHawkes has multiple potential explanations. A first hint, as explained in [93], is the presence of larger inter-event durations than can be accounted for by the Hawkes model. Next, in financial markets, it is clear that many positive feedback mechanisms are present at different time scales (see [90] for a list). These mechanisms range from reactive high-frequency trading at short time scales to herd behaviour of traders at long time scales. Thus a more complete model would consider self-excitation at multiple time scales (and magnitudes), as well as their interactions. RHawkes constitutes a simple instance that allows for first order dependency between clusters.

2.7 Discussion

We have proposed a new type of self-excited point process that extends the Hawkes model to allow for the immigration process (which is a homogeneous or IPP within the standard Hawkes framework) to

be a stochastic self-excited process as well. Within the specific model discussed in the paper (RHawkes: Hawkes process with renewal immigration), the immigration process was considered to be a renewal process with a given intensity function. However, this generalization makes direct MLE impossible.

We have made the estimation of RHawkes possible by the introduction of two EM algorithms: EM1, which is built on the existing EM algorithm for Hawkes processes (EM0), and EM2 which uses a reduced set of missing information. These estimation techniques were shown to be consistent in simulation studies, and easily allow for non-parametric estimation. Further, we have shown that the EM2 algorithm can be applied to the estimation of the Hawkes process with IPP immigration, and this application is more computationally efficient than the standard (EM0) approach.

The (standard) Hawkes model is widely used in many areas from physics to quantitative finance. In such applications, in particular when the focus is the quantification of the branching ratio, we recommend as a best practice to consider both IPP and renewal immigration as alternative models.

We have provided an example of the relevance of our results to the existing literature on the quantification of the branching ratio for high-frequency price fluctuations (see for example [91, 112]). In the provided case study, it was shown that the RHawkes model is superior to the specifications considered in [91], thus questioning the validity of the Poisson immigration assumption in such applications. As another example, for the modeling of rainfall, our approach provides a richer model than [58] and theoretically superior estimation to [219]. Finally, extending EM1 and EM2 for the estimation of spatio-temporal and marked versions of the Hawkes model would allow one to test the validity of the Poisson background intensity hypothesis in modeling triggered seismicity.

Chapter 3

The ARMA point process

This chapter is part of a work-in-progress.

3.1 Introduction

I argue that, by driving the Hawkes (autoregressive) process (2.1) with NS (moving average) process (1.3) immigration, one obtains a process that, when aggregated, is approximately the INARMA process (the ARMA time series for positive integers). We thus propose that this process be called the ARMA point process. A special case of this process, called the *dynamic contagion* process – having exponentially decaying triggering intensities and exponential marks – was introduced in [66]. Here an emphasis is placed on estimation of the model.

As is the general case, once a moving average component is included in a model, the estimation – namely maximum likelihood (MLE) – becomes more difficult. In the time series case this is because the innovations are not observed, and in the point process case, because the branching structure is not observed. For instance, in [36] the INMA process is estimated by conditional least squares or generalized method of moments. For NS processes one often uses method of moments or maximizes a quasi (palm) likelihood [272, 205]. In [65] such processes have been estimated by filtering the (unobserved) intensity under a Gaussian process approximation. For the INARMA models a Bayesian MCMC technique was used [188, 83]. As suggested in [38] a frequency domain estimator that approximates MLE for point process models – including the Hawkes and NS processes [118], NS processes with renewal immigration [209, 46] and INAR models [231] – has been used. Here an Expectation Maximization (EM) algorithm is defined for the estimation of the ARMA point process, thus containing an algorithm for the Hawkes

process, the NS process and their INARMA analogies. Efficient simulation is also addressed.

3.2 The ARMA Point Process

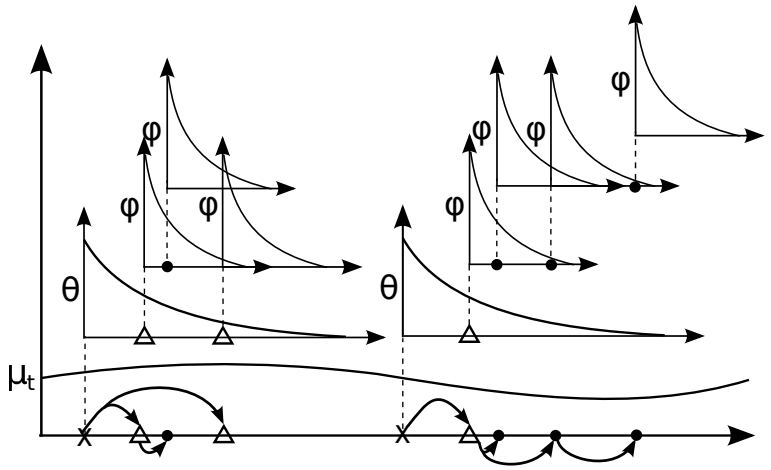

Figure 3.1: A realization of the ARMA point process with innovation/immigration intensity μ_t, MA intensity θ, and AR intensity ϕ. Immigrants, θ-offspring and ϕ-offspring are denoted by x mark, triangle, and dot, respectively. A point is connected to the intensity that it triggers by a vertical dashed line. All points are projected onto the horizontal axis, with parenthood indicated by arrows, forming the full realization.

Here we introduce the ARMA point process, defined by its *CIF* (conditional intensity function),

$$\begin{aligned}
\lambda(t|\mathcal{F}_t^Z) &= \lim_{\Delta\downarrow 0} \Delta^{-1}\mathrm{E}\left[N(t,t+\Delta)|\mathcal{F}_t^Z\right] \\
&= \mu + \int_{-\infty}^{t} \theta(t-s)dN^{\mu}(s) + \int_{-\infty}^{t} \phi(t-s)dN(s) \\
&= \mu + \sum_{j=1}^{N(t)} Z_j^{\mu}\theta(t-t_j) + \sum_{k=1}^{N(t)} \phi(t-t_k) \ , \qquad (3.1)
\end{aligned}$$

which drives the Hawkes process (1.4) with shot noise (1.3). The process is conditioned on the natural filtration $\mathcal{F}_t^Z$ that contains the entire history of the process, as well as the indicator variables: for $i \in \mathbb{Z}$, $Z_i^{\mu} = 1$ if t_i is an immigrant and $Z_i^{\mu} = 0$ otherwise. That is, $Z_i^{\mu} = dN^{\mu}(t_i)$ where $N^{\mu}(s) := \int_0^s dN^{\mu}(s)ds$ is the immigrant counting process. Here $\mu > 0$ is the immigration/innovation intensity, ϕ the AR intensity, and θ the MA intensity. The indicator variable $Z_i^{\mu} = 1$ if t_i is an immigrant (i.e., when $dN^{\mu}(t_i) = 1$), and $Z_i^{\mu} = 0$ otherwise. Thus the MA intensity is only triggered by immigrants. The MA and AR intensities are intensities of IPP, and can thus be factored into pdf and non-negative normalizing constants, $\theta(\cdot) = \gamma g(\cdot)$ and $\phi(\cdot) = \eta f(\cdot)$, as defined in the NS (1.3) and Hawkes processes

(1.4). Taking the unconditional expectation of the CIF (3.1) yields the expected intensity,

$$\bar{\lambda} = \frac{\mu(1+\gamma)}{1-\eta} \ , \tag{3.2}$$

indicating that $\eta < \infty$ and $\gamma < 1$ are necessary conditions for stationarity, and $\eta > 1$ will produce an explosive proliferation of points. By setting $\gamma = 0$ one recovers the Hawkes process, and by setting $\eta = 0$ one recovers the NS process. Formulating the ARMA point process as a branching process brings great insight, as visualised in Fig. 3.1: μ introduces *immigrants*, which may then trigger a single generation of θ-*offspring* with intensity $\theta(\cdot)$, and then all existing points trigger a generation of ϕ-*offspring* with intensity $\phi(\cdot)$, which may, in turn, trigger the subsequent generation of ϕ-*offspring* in the same way. The sum of these independent IPP provides the ARMA CIF (3.1), and the set of immigrant and (θ- and ϕ-)offspring points forms the ARMA point process realization. The MA *branching ratio* γ is the expected number of θ-*offspring* of a single immigrant, and the AR *branching ratio* η is the expected number of immediate ϕ-*offspring* of any point. Further, counting all generations, a single point is expected to produce $\eta + \eta^2 + \cdots = 1 - 1/(1-\eta)$ ϕ-*offspring*. Thus, as in the Hawkes process, η is the expected proportion of all points that are ϕ-*offspring*.

3.3 The Relationship to Integer ARMA Models

The discrete valued analogues of classical time series models [39] have seen a flurry of recent development [97, 181] and enjoy many current and potential applications. Here we consider the INARMA(p, q) process, an ARMA process for counts $X_l \in \{0, 1, 2, \dots\}$, $l \in \mathbb{Z}$, that satisfies the difference equation,

$$X_l = \epsilon_l + \sum_{k=1}^{q} \widetilde{\theta}_k \circ \epsilon_{l-k} + \sum_{j=1}^{p} \widetilde{\phi}_j \circ X_{l-j} \ , \ \epsilon_l \overset{i.i.d}{\sim} \text{Poisson}(\widetilde{\mu}) \ , \ \widetilde{\mu} \geq 0 \tag{3.3}$$

which has the familiar ARMA structure, but with a thinning operator, instead of multiplication, to preserve the count value of the process. Here the *Poisson thinning* operator $\circ$, for some count variable Z is considered,

$$\alpha \circ Z = \sum_{i=1}^{Z} Y_i, \ \alpha > 0, \text{ and } \alpha \circ 0 := 0 \ , \tag{3.4}$$

where all (Y_i) and Z are independent. If $Y_i \overset{i.i.d}{\sim} \text{Pois}(\alpha)$, then $\alpha \circ Z | Z = z$ is a sum of z independent Poisson variables with parameter α, and thus has distribution $\text{Pois}(\alpha z)$. Thus, given that both all

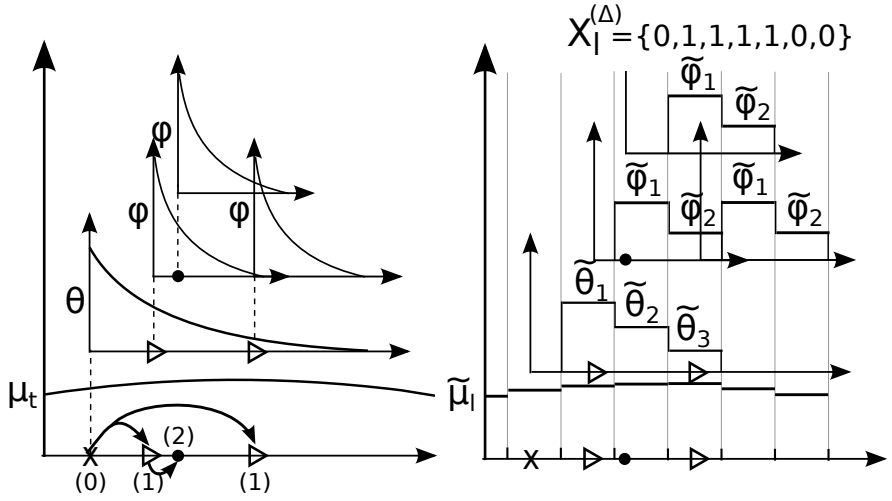

Figure 3.2: Identical realisations from (left plot) the ARMA point process (3.1), and (right plot) the INARMA$(2,3)$ process. The left plot is the first cluster from fig. 3.1, where the symbols are defined. The right plot is given for grid Δ (3.3) indicated by the vertical lines, where the count values are given above the plot. In the INARMA process one knows the counts, not the exact locations of the points in time. Here the points are included to show how the INARMA process approximates the Hawkes process. The origins of the axes framing the AR and MA triggering coefficients are located at the time values of the points that triggered them to highlight that the INARMA process cannot trigger points within the bin containing the point that generated the triggering function.

thinnings in (3.3) are independent of eachother, and of the Poisson innovation ϵ_l, then the conditional df of $X_l|X_{(l-p):(l-1)}, \epsilon_{(l-q):(l-1)}$ is also Poisson. The unconditional df of X_l is not Poisson. It is important to note that the standard thinning used in integer time series is Bernoulli/Binomial thinning, where the variable Y has a Bernoulli df. In this case, the unconditional df of X_l is Poisson, but the conditional one is not. A survey of the different thinning specifications employed within the literature are summarized in [286]. Bernoulli thinning provides the most mathematical elegance, and maintains many structural similarities to ARMA models. Other forms of thinning, such as Poisson thinning, provide useful generalisations – especially when the unconfiditional df is not Poisson. Specifically, Poisson thinning is used here to enable the connection with the ARMA point process. In this sense, each model motivates and justifies the other.

The INARMA process (3.3) is a multi-type branching process [72] that is stationary for $\sum_{j=1}^{p} \widetilde{\phi}_j < 1$ and $\sum_{k=1}^{q} \widetilde{\theta}_k < \infty$. The branching interpretation is the same as for the ARMA point process: the innovation count ϵ_l introduces immigrants, and thinning (3.4) has the interpretation that each of the X_j points in the j^{th} bin is expected to produce $\widetilde{\phi}_{l-j}$ offspring in the l^{th} bin, where $l > j$. Thus the INARMA process introduces a burst of offspring triggered by immigrants, via thinning with $\widetilde{\theta}$

coefficents, and an autoregressive tree of offspring triggered by all past events, via thinning with $\widetilde{\phi}$ coefficients. The INARMA process is more general in the sense that, for negative coefficients $\widetilde{\theta}$ and $\widetilde{\phi}$, it can produce "self-inhibiting" behaviour.

Next a formal argument is made to demonstrate an asymptotic equivalence between these two models. This is aided by observing the similarities between the two subplots of Fig. 3.2. The INARMA(p, q) process, with Poisson thinning, has conditional expectation,

$$E[X_l|X_{(l-1):(l-p)}, \epsilon_{(l-1):(l-q)}] = \widetilde{\mu} + \sum_{k=1}^{q} \widetilde{\theta}_k \epsilon_{l-k} + \sum_{j=1}^{p} \widetilde{\phi}_j X_{l-j} \ , \tag{3.5}$$

and therefore, dividing (3.5) by Δ, one obtains a discrete CIF of the INARMA process.

Next, if one aggregates the ARMA point process (3.1) on bins of width $\Delta > 0$, one obtains the counting variables $\{X_l^{(\Delta)} = N\left(\Delta l, \Delta(l+1)\right),\ l = 1, ..., \frac{r}{\Delta}\}$ for all points, and $\{\epsilon_l^{(\Delta)} = N^{\mu}\left(\Delta l, \Delta(l+1)\right),\ l = 1, ..., \frac{r}{\Delta}\}$ for innovations, where $N^{\mu}(s) = \int_0^s dN^{\mu}(s)$ only counts immigrants, as in the NS process (1.3). For instance, in Fig. 3.2, the immigrant counts are: $\{\epsilon^{(\Delta)}\} = \{0, 1, 0, 0, \dots\}$. In this aggregated model, one obtains the discrete CIF,

$$\Delta^{-1}\mathrm{E}\left[X_l^{(\Delta)}|X_{(l-1):(l-p)}^{(\Delta)}, \epsilon_{(l-1):(l-q)}^{(\Delta)}\right] \approx \mu + \sum_{k=1}^{q} \theta(k\Delta)\epsilon_{l-k}^{(\Delta)} + \sum_{j=1}^{p} \phi(j\Delta)X_{l-j}^{(\Delta)} \ , \tag{3.6}$$

where the approximation treats ϕ and θ as step functions, with step width Δ, and (falsely) excludes triggering of points within the same bin. Roughly speaking, the approximation becomes exact as $\Delta \to 0^+$ and $p, q \to \infty$. In this limit the conditional intensities of (3.5) and (3.6) become equivalent by equating parameters: $\widetilde{\mu} = \mu\Delta$, $\widetilde{\theta_k} = \Delta\theta(k\Delta)$, and $\widetilde{\phi_j} = \Delta\phi(j\Delta)$. Since the CIF uniquely characterizes the finite-dimensional distributions of point processes [62], the processes are thus equivalent.

As special cases, the Hawkes process is approximated by the INAR process, and the NS process is approximated by an INMA process. The (weak) converge of the INAR process to the Hawkes process, as $\Delta \to 0$, was rigorously proved in [150]. The INAR approximation was exploited in [149] to approximately estimate the Hawkes process using a non-parametric least-squares-based estimator.

As is the case for the ARMA point process, the INARMA model cannot be directly/simultaneously estimated by MLE due to missing information – here being the innovation counts $\epsilon_{l-1}^{(\Delta)}, \epsilon_{l-2}^{(\Delta)}, \dots$ where only the complete counts $X_{l-1}^{(\Delta)}, X_{l-2}^{(\Delta)}, \dots$ are observed. In Sec. 3.5 we provide an EM algorithm to estimate the ARMA point process, which may also be applied to the INARMA model.

3.3.1 Further details about the INARMA model

The thinned variable is a compound random variable. One can use the Tower property of conditional expectations to derive the moments of such variables. For instance, for Poisson thinning, where Z is a count random variable, and $\alpha, \beta \geq 0$: $E[\alpha \circ Z] = \alpha E[Z]$, $E[\beta \circ \alpha \circ Z] = \alpha\beta E[Z]$, and $E[(\alpha \circ Z)(\beta \circ Z)] = \alpha\beta E[Z^2]$. Thus such moments are the same if one replaces the thinning with multiplication. However the Poisson thinning case differs from the multiplicative case when the second moment is considered, as $E[(\alpha \circ Z)^2] = \alpha E[Z] + \alpha^2 E[Z^2]$. Further, $E[(\beta \circ \alpha \circ Z)(\alpha \circ Z)] = \beta E[(\alpha \circ Z)^2]$.

To consider simple models, for the INMA(1) process with Poisson thinning,

$$\mathrm{E}[X] = (1 + \theta_1)\mathrm{E}[\epsilon] \ , \tag{3.7}$$

$$\mathrm{Var}(X) = (1 + \theta_1^2)\mathrm{Var}(\epsilon) + \theta_1 \mathrm{E}[\epsilon] \ , \tag{3.8}$$

$$\mathrm{Cov}(X_l, X_{l-1}) = \theta_1 \mathrm{Var}(\epsilon) \ , \tag{3.9}$$

and thus the mean and covariance are the same as the standard MA(1) process, but the variance is different, due to the presence of the $\theta_1 \mathrm{E}[\epsilon]$ term. For the INAR(1) process with Poisson thinning,

$$\mathrm{E}[X] = (1 - \phi_1)^{-1}\mathrm{E}[\epsilon] \ , \tag{3.10}$$

$$\mathrm{Var}(X) = (1 - \phi_1^2)^{-1}(\mathrm{Var}(\epsilon) + \phi_1 \mathrm{E}[X]) \ , \tag{3.11}$$

$$\mathrm{Cov}(X_l, X_{l-1}) = \phi_1 \mathrm{Var}(X) \ , \tag{3.12}$$

and thus the mean is the same as the standard AR(1) process, but the variance is different, due to the presence of the $\phi_1 \mathrm{E}[X]$ term. Thus, the covariance formula, as written above, is the same as the standard AR(1) process, but, for equal parameters, will have a different autocovariance, due to the different variance.

Next, the Yule-Walker approach for systematically computing autocorrelations for the INARMA(p, q) process is shown. Take the INARMA(p, q) process,

$$X_l = \epsilon_l + \sum_{k=1}^{q} \theta_k \circ \epsilon_{l-k} + \sum_{j=0}^{p} \phi_j \circ X_{l-j} \ . \tag{3.13}$$

For $u = 0, 1, 2, \ldots$ and $i \in \mathbb{Z}$, denote $a_u = E[X_l X_{l-u}]$ $(= E[X_{l+i}\epsilon_{l+i-u}])$ and $b_u = E[X_l \epsilon_{l-u}]$ $(=$

$E[X_{l+i}\epsilon_{l+i-u}]$), so $b_u = 0$ for $u < 0$. Then, multiplying (3.13) by ϵ_{l-u}, and taking the expectation, gives the system of equations,

$$b_u = \sum_{k=0}^{q} \theta_k E[\epsilon_{l-k}\epsilon_{l-u}] - \sum_{j=1}^{q} \phi_j b_{u-j} \ , \tag{3.14}$$

where $\theta_0 = 1$, $b_{u-j} = 0$ for $u < j$, and $b_0 = E[\epsilon_l^2] + E[\epsilon_l]E[X]$. Thus $b_1, \ldots, b_q$ may be determined recursively using eq. 3.14. Next, multiplying (3.13) by X_{l-u}, and taking the expectation, gives the system of equations,

$$\sum_{j=0}^{p} \phi_j a_{u-j} = \sum_{k=1}^{q} \theta_k b_{k-u} \ , \tag{3.15}$$

where $\phi_0 = 1$, and given $b_1, \ldots, b_q$, one can solve for $a_1, \ldots, a_q$, and $a_{q+1}, a_{q+2}, \ldots$ are solved where the right hand side of (3.15) is zero.

3.4 Simulation of the ARMA point process

Below an algorithm for simulating ARMA point processes is presented. It exploits the fact that, given the branching structure, the innovation, MA, and AR processes are mutually independent IPP. At the end of step II one has simulated an NS process (1.3). By skipping step II, and completing step III, one simulates a Hawkes process (1.4). To avoid edge effects, one should simulate on a large window, and discard the *burn in period*. Using inverse transform sampling makes the algorithm very fast. For instance, simulating with $\gamma = 1$, and $\eta = 0.9$, $O(10^4)$ points can be simulated in $O(0.1)$ seconds (implemented in R on a standard laptop with 2.90 GHz processor, and 4GB ram). Simulation algorithms for the Hawkes process follow the same approach, but their implementations have been much slower due to using inefficient IPP sampling techniques [184, 192], rather than using inverse transform sampling, as discussed in Chap. 1.

Simulation algorithm

I. Simulate the immigrant points

(i) Simulate Poisson process $\{T_i^{(0)}\}_{i\in 1,\ldots,n_0}$ on the window $(0,r]$, where $N(r)=n_0$.

II. Simulate MA points

(i) For each immigrant $i=1:n_0$: simulate the number of offspring $N_i^\theta \overset{i.i.d}{\sim} \text{Pois}(\gamma)$, and then sample the N_i^θ inter-event times $S_{i,j},\ j=1:N_i^\theta$, i.i.d from pdf g. If $N_i^\theta=0$, simulate zero inter-event times for that immigrant.

(ii) The MA points generated by the i^{th} immigrant are then $\{T_{i,j}^\theta\}_{j=0:N_i^\theta} = T_i^{(0)} + \{S_{i,j}\}_{j=0:N_i^\theta}$.

(iii) The immigrant and MA points together are $\{T_i^{(1)}\}_{i\in 1,\ldots,n_1} = \{T_i^{(0)}\} \cup \{T_{1,j}^\theta\} \cup \cdots \cup \{T_{n_0,j}^\theta\}$, where $n_1 = n_0 + \sum_{i=1}^{n_0} N_i^\theta$.

III. Simulate AR points by generation

(i) Set the fertile points $A=\{1,\ldots,n_1\}$, generation $k=1$, and zeroeth generation points $\{T_i^{\phi[0]}\} = \{T_i^{(1)}\}$.

(ii) For the current generation k, for all fertile points $\forall i \in A$: simulate the number of direct offspring $N_i^{\phi[k]} \overset{i.i.d}{\sim} \text{Pois}(\eta)$, and then the $N_i^{\phi[k]}$ inter-event times $S_{i,j}^{\phi[k]},\ j=1,\ldots,N_i^{\phi[k]}$, i.i.d from pdf f. If $N_i^{\phi[k]}=0$, simulate zero inter-event times for that that point.

(iii) The AR points generated by the i^th point are then $\{T_{i,j}^{\phi[k]}\}_{j=0:N_i^{\phi[k]}} = T_i^{\phi[k-1]} + \{S_{i,j}^{\phi[k]}\}_{j=0:N_i^{\phi[k]}}$ and

(iv) the union of these sets, $\{T^{\phi[k]}\} = \{T_{1,j}^{\phi[k]}\} \cup \cdots \cup \{T_{N_i^{\phi[k]},j}^{\phi[k]}\}$, is the offspring of generation k.

(v) Update the fertile set $A=\{i : T_i^{\phi[k]} < r\}$ to be all points born in the current generation k that fall within the window $(0,r]$.

(vi) If A is non-empty then increment the generation ($k=k+1$) and return to (ii), otherwise return the realization formed by joining all generations: $\{T_i\}_{i=1:n} = \{T_i^{(1)}\} \cup \{T^{\phi[1]}\} \cup \cdots \cup \{T^{\phi[k]}\}$.

3.5 EM algorithm for the estimation of the ARMA point process

Here the observed data is $\boldsymbol{X} = \boldsymbol{t}_{1:n}$ and the missing data is indicator variables $\boldsymbol{Z} = \{Z_i^\mu, Z_{i,j}^\theta, Z_{i,j}^\phi,\ i = 1, ..., n,\ j = 1, ..., i-1\}$, which are zero except $Z_i^\mu = 1$ if t_i is an immigrant, $Z_{i,j}^\theta = 1$ if t_i is triggered by $\theta(t - t_j)$, and $Z_{i,j}^\phi = 1$ if t_i is triggered by $\phi(t - t_j)$. Thus, $\forall i$: $Z_i^\mu + \sum_{j=1}^{i-1}(Z_{i,j}^\theta + Z_{i,j}^\phi) = 1$. Given the missing data, the first term of the likelihood (1.9) is decomposed as,

$$\prod_{i=1}^{n} \lambda(t_i|\boldsymbol{t}_{1:n}, \boldsymbol{Z}) = \prod_{i=1}^{n} \Big[\ \mu^{Z_i^\mu} \prod_{j=1}^{i-1} \Big[\theta(t_i - t_j)\Big]^{Z_{i,j}^\theta} \prod_{k=1}^{i-1} \Big[\phi(t_i - t_k)\Big]^{Z_{i,k}^\phi}\ \Big]\ , \tag{3.16}$$

where an intensity is only evaluated at the correct time or inter-event time as determined by the indicator variables that encode the branching structure. Taking the log and the expectation of the likelihood, the *Expected Complete Loglikelihood* (abbreviated as Q) is,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z}, \hat{\boldsymbol{\beta}}^{[m]}) &= \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{t}_{1:n}, \hat{\boldsymbol{\beta}}^{[m]}}\Big[\log \mathrm{L}(\boldsymbol{\beta}|\boldsymbol{t}_{1:n}, \boldsymbol{Z})\Big] = Q_\mu + Q_\theta + Q_\phi\ , \\ Q_\mu &= \sum_{i=1}^{n} \pi_i^\mu \log\mu - \int_0^T \mu ds\ , \\ Q_\theta &= \sum_{i=1}^{n}\sum_{j=1}^{i-1} \pi_{i,j}^\theta \log\theta(t_i - t_j) - \sum_{j=1}^{n} \pi_j^\mu \int_{t_j}^{r} \theta(s - t_j) ds\ , \\ Q_\phi &= \sum_{i=1}^{n}\sum_{k=1}^{i-1} \pi_{i,j}^\phi \log\phi(t_i - t_k) - \sum_{k=1}^{n} \int_{t_k}^{r} \phi(s - t_k) ds\ , \end{aligned} \tag{3.17}$$

where innovation, MA, and AR processes are decoupled into Q_μ, Q_θ, and Q_ϕ respectively. The probability weights in (3.17), to be calculated in the E step, are given by,

$$\pi_i^\mu = \Pr\{Z_i^\mu = 1|\boldsymbol{t}_{1:n}, \boldsymbol{\beta}\} = \frac{\mu_{t_i}}{\mu_{t_i} + \sum_{j=1}^{i-1} \pi_j^\mu \theta(t_i - t_j) + \sum_{j=1}^{i-1} \phi(t_i - t_j)} \tag{3.18}$$

$$\pi_{i,j}^\theta = \Pr\{Z_{i,j}^\theta = 1|\boldsymbol{t}_{1:n}, \boldsymbol{\beta}\} = \frac{\pi_j^\mu \theta(t_i - t_j)}{\mu + \sum_{j=1}^{i-1} \pi_j^\mu \theta(t_i - t_j) + \sum_{j=1}^{i-1} \phi(t_i - t_j)} \tag{3.19}$$

$$\pi_{i,j}^\phi = \Pr\{Z_{i,j}^\phi = 1|\boldsymbol{t}_{1:n}, \boldsymbol{\beta}\} = \frac{\phi(t_i - t_j)}{\mu_{t_i} + \sum_{j=1}^{i-1} \pi_j^\mu \theta(t_i - t_j) + \sum_{j=1}^{i-1} \phi(t_i - t_j)}\ , \tag{3.20}$$

exploiting the thinning property [62] whereby the probability that t_i comes from one of the independent (sub-)processes is equal to that processes' share of the total CIF at t_i.

Next, considering the M-step, Q_μ, Q_θ, and Q_ϕ (3.17) have the structure of log-likelihoods for density estimation with sample weights, where the samples are all positive interevent times $t_i - t_j,\ j <$

$i = 1, \ldots, n$, and the weights are the prefactors of the log terms in (3.17). In further detail: Q_μ has a simplified structure due to being a Poisson process and has estimator,

$$\hat{\mu} = \sum_{i=1}^{n} \pi_i^\mu / r \ , \tag{3.21}$$

in the homogeneous case, and may easily be extended to the inhomogeneous case with by estimating a density for points $t_{1:n}$ with weights $\pi_{1:n}^\mu$, and re-scaling the density to have mass $\sum_{i=1}^n \pi_i^\mu$. For Q_θ and Q_ϕ, by factoring the intensities into their branching ratios and densities – $\theta(\cdot) = \gamma g(\cdot)$ and $\phi(\cdot) = \eta f(\cdot)$ – the densities and branching ratios may be estimated separately. As an example, for $g(\cdot)$ being an exponential density, its scale estimate is the weighted average,

$$\hat{\tau} = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \pi_{i,j}^\theta \cdot (t_i - t_j) \Big/ \sum_{i=1}^{n} \sum_{j=1}^{i-1} \pi_{i,j}^\theta. \tag{3.22}$$

Finally, the MLE for the branching ratios are,

$$\widehat{\gamma} = \frac{\sum_{i=1}^n \sum_{j=1}^{i-1} \pi_{i,j}^\theta}{\sum_{i=1}^n \pi_i^\mu \widehat{G}(r - t_i)} \quad , \quad \widehat{\eta} = \frac{\sum_{i=1}^n \sum_{j=1}^{i-1} \pi_{i,j}^\phi}{\sum_{i=1}^n \widehat{F}(r - t_i)} \ , \tag{3.23}$$

where G and F are the CDF of g and f respectively, and their role in the denominator is to account for offspring truncated by the end of the observation window r.

This algorithm can be adapted to INARMA (3.3) estimation by defining the df to be histograms with bin width Δ equal to the time between the INARMA counting variables. The algorithm requires storage and computation with matrices that are $O(n^2)$. On a standard PC this makes computation prohibitive for samples with $n > 10^4$. However this implementation is crude as, in this case, even the largest interevent time $t_n - t_1$ is considered as an interevent time by which t_n could be triggered via $\theta(t - t_1)$ or $\phi(t - t_1)$, despite the fact that the probablity of this could be effectively 0. Thus, for window r large relative to the support of θ and ϕ, one can safely omit interevent times above a certain threshold. The result of this is banded/sparse matrices which reduce storage and computation from n^2 to $n \times m$, potentially with $m << n$. This enables fast estimation on, e.g., samples with $n = O(10^5)$ with memory/point interaction limited to a neighbourhood of $m < 1000$ adjacent points.

Regarding the convergence of the EM algorithm [218, 295, 71], one first needs that the necessary MLE regularity conditions are satisfied [191], for instance having smooth distributions that are not

too heavy tailed. Next, it must be ensured that the sequence of parameter estimates does not reach the boundary of the parameter space. For instance, if estimates $\mu^{[m]}$, $\eta^{[m]}$, or $\gamma^{[m]}$ are equal to zero at any iteration, m, or equivalently, if the support of $f^{[m]}(.)$ or $g^{[m]}(.)$ is smaller than the smallest interevent time, then the estimates will remain zero (eqs. 3.18-3.20). However, given non-zero starting estimates, the EM algorithm estimates automatically satisfy the constraints of the model parameters. Thus one simply needs to avoid problematic initial estimates. Regarding speed of convergence, there is the general result of [71, 295] that the algorithm will not worsen the likelihood with each iteration. Further, from [218], given that Q is differentiable in Θ and the M-step has a unique solution, then the EM algorithm iterates in a positive direction on the true likelihood surface. Finally, when the missing information is small compared to the complete information, EM exhibits approximate Newton behavior with superlinear convergence near the true optimum. In terms of the ARMA point process, as well as other mixture type models, this means that when clusters are overlapping (missing information is large) convergence will be slow, as has been shown for the Hawkes process with exponential offspring df [164, 260], as well as other mixture models [297, 218]. Exactly deriving the convergence properties of the ARMA point process model would add no new qualitative insight, and would lack generality due to only applying for a single parameterization.

3.6 Discussion

The aggregated NS/shot noise process is approximately an INMA process, the aggregated Hawkes process is approximately an INAR process, and the aggregated Hawkes process driven by shot noise (the ARMA point process, or dynamic contageon process) is approximately an INARMA process. As the aggregation becomes fine, the integer time series becomes equivalent to the analogous continuous time point process. The univariate unmarked ARMA point process was considered but multivariate and marked extensions follow immediately. The EM algorithm for the ARMA point process provides MLE for this new type of process, as well as the NS process that it contains. Further, the algorithm can be generalized to the aggregated INARMA case. This provides MLE to a class of models where many inferior estimation algorithms had been used (See Sec. 1).

Much basic work on the development of this model remains. For instance, the second order statistics need to be derived for the ARMA point process. Deriving the MA (and AR) representations for the INARMA model – as done in [150] for the INAR case – is related to this. Further, the limiting equivalence between the INARMA model and the ARMA point process should be rigorously proven as has been done for the Hawkes process and the INAR model [150], or by using probability generating functionals. The performance of the EM algorithm for the estimation of the ARMA point processes should be characterized, and compared with other estimators, in simulation studies.

Part II

A general outlier test, & singular "dragon-king" extremes

This short chapter is an extended version of a piece written by Spencer Wheatley and Didier Sornette to appear in an edition of SATW INFO edited by Prof. Wolfgang Kröger.

Chapter 4

Dragon-kings and extremes

Introduction

Extremes dominate the long term quality and organization of most important natural and societal systems: the largest two nuclear power plant accidents have caused five times more damage than all other (>200) historical accidents together [292]; the largest ten percent of private data breaches from organizations accounts for ninety-nine percent of the total breached private information[289]; the largest five epidemics since 1900 caused twenty times the fatalities of all others [290, 106]; etc. Such statistics are consistent with (extremely) heavy tailed distributions. Furthermore, it is often the case that the largest events are outliers and have special circumstances associated with them. This point will be elaborated below.

Despite the importance of extreme events, due to ignorance, misaligned incentives, and cognitive biases, we often fail to adequately anticipate them. Technically speaking, we choose models that are not heavy-tailed enough, and under-appreciate both serial and multivariate dependence of extremes. Some examples of such failures in risk assessment include: the use of Gaussian models in finance (e.g., Guassian copula and Black-Scholes models); the use of Gaussian processes and linear wave theory failing to predict the occurrence of rogue waves; the failure of economic models to predict the 2007 financial crisis; and the under-appreciation of external events, cascades, and nonlinear effects in probabilistic risk assessment. Such high impact failures (e.g., the 2011 Fukushima disaster) emphasize the importance of the study of extremes.

Dragon-kings & Black Swans

Here a special type of "outlying" extreme event is discussed. Dragon-king (DK) is a double metaphor for an event that is both extremely large in size or impact (the "king" is the richest in the land) and generated from a unique process/origin (a "dragon") relative to other events from the same system. This term was introduced by Prof. Didier Sornette. As an example, a king may himself be a dragon-king in terms of wealth within his country, as he has the unique right to tax the population. DK are generated by mechanisms such as positive feedback, tipping points, bifurcations, and phase transitions, that tend to occur in nonlinear and complex systems, and amplify DK events to disproportionately extreme levels. By understanding and monitoring these dynamics, some predictability of such events may be obtained [259, 246, 241]. Predictability aside, it is also of importance to i) identify the possibility, and ii) assess the risk of such events. The statistics and mechanisms of DK are further elaborated below.

It is worth contrasting the DK of Sornette with the Black swan theory popularized by Taleb [271], which is related to well established concepts such as Knightian uncertainty [152] and the sampling of species problem [298]. Black swan is a metaphor for an event that is surprising (to the observer), has a major effect/impact, and whose occurence is rationalized in hindsight [271]. An analysis of the meaning of the concept in a risk context was given by Aven [20]. Taleb claims that black swan events are not predictable, and in practice encourages one to prepare rather than predict, and limit ones exposure to extreme fluctuations. However, in a wide range of physical systems, many extreme events are predictable to some degree, provided that one has a sufficiently deep understanding of the structure and dynamics of the focal system, and the ability to monitor it [13, 241, 246]. Thus, many extremes are DK rather than black swans, and the distinction is important.

Dragon-king mythology

Here the Dragon-King terminology is motivated and explained as it relates to a symbol common in a variety of ancient mythologies. In fig. 4.1 an illustration of the Ouroboros character is given ("he that eats his tail" in Greek), reproduced from [77]. This ancient character is a symbol of alchemy, or more generally a reflexive or self-referential process. That the character can be represented by a dragon and with a king's crown is very appropriate. The structure of the Ouroboros can also be

Figure 4.1: The Ouroboros character. This illustration was taken from [77]

related to positive feedback, unsustainable self-consumption (creation arising from destruction), and other features of DK events. Further, the Ouroboros and other dragon/serpent characters are related to extreme events in multiple mythologies: e.g., the Leviathan of the Old Testament, Jörmungandr of Norse mythology, and the Dragon-King of Chinese mythology.

Statistics of extremes & beyond

Many phenomena in both the natural and social sciences have power law statistics [183, 189, 243]. Furthermore Extreme Value Theory (EVT) provides that many distributions (the Frechet class) have tails that are asymptotically power law [79] with the Generalized Pareto Distribution (GPD). The result of this is that, when dealing with crises and extremes, power law tails are the "normal" case. Power laws have a unique property (scale-invariance/self-similarity), implying that all events – both large and small – are generated by the same mechanism, and thus there will be no distinct precursors by which the largest events may be predicted.

However, in a variety of studies it has been found that, despite the fact that a power law models the tail of the empirical distribution well, the largest events are significantly outlying [290, 142, 201]. Such events are interpreted as DKs when they indicate a departure from the generic process underlying the power law. An instance of this is given in Figs. 4.2 and 4.3. Examples of where outliers have been found include crashes in financial markets, the cost and radiation released in nuclear power plant accidents, urban agglomeration sizes within a country [290], intraday wholesale electricity prices [138], the 2007 stock market crash [212] etc. That is, DK are statistical outliers that are highly informative, and should be the focus of much statistical attention. Thanks to a key result from EVT, a quite general

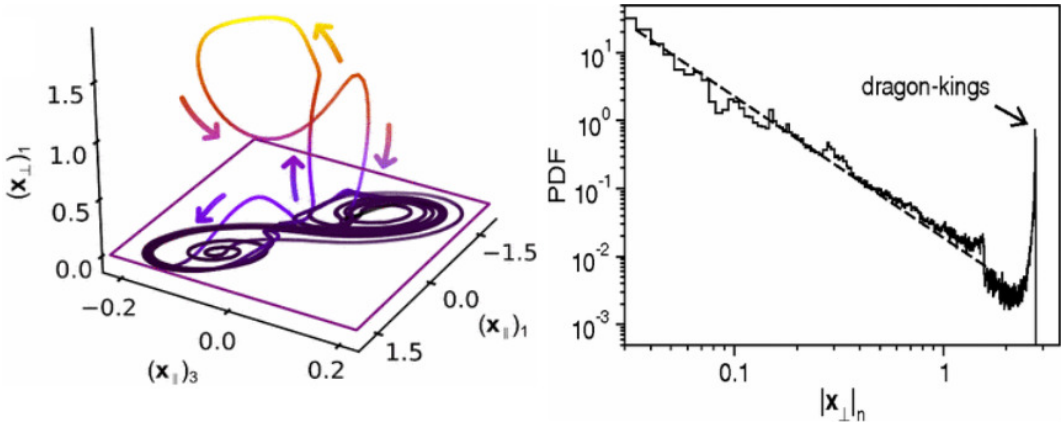

Figure 4.2: Left: Illustration of the system trajectory in the vicinity of a bubbling event. Right: empirical probability density function (histogram) of peak heights in trajectories in double logarithmic scale. This illustration was taken from [45]

outlier test is available for detecting DK and assessing their significance [290].

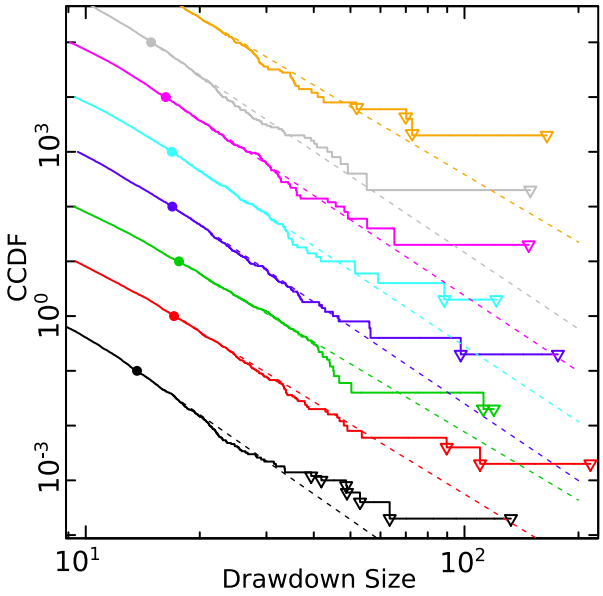

Figure 4.3: The 5000 largest drawdowns (a measure of crashes) for 8 different futures contracts plotted according to their empirical complementary cumulative distribution function (CCDF), shifted by factors of 10 for visibility. The dashed lines are power law fits, and the triangles the detected drawdowns [290].

On the topic of EVT, there is another important point to make: the distinction between the statements (i) "a structural break / regime change produces the events that are outlying (e.g., with

respect to EVT)", and (ii) "the impact of the outlying event leads to the structural break / regime change". The distinction is important because extreme events often occur around regime changes, and the first statement could attribute the outlying event to this change in regime, while the second would suggest the opposite causal relationship. The first statement may be used as an explanation for the failure of EVT when some change – typically treated as exogenous – takes place. That the change/shock is exogenous agrees with the black swan concept. At the same time, a black swan event is defined as having extreme consequences, e.g., causing a regime change, rather than itself being caused by one. The second statement is thus compatible with the black swan and DK events.

On the other hand, the apparent discrepancy between the two statements may falsely rely on the conflation of the "regime change", that is the DK mechanism itself, with the regime change that follows the DK event; e.g., a reaction to the crisis induced by the DK. Further, events with outlying/oversized impact are likely to lead to regime changes, by definition. Thus, if one accepts that there are extreme events that are beyond/outlying relative to some model, e.g,. an EVT model, then the important question is: are they treated as exogenous, or can they be "endogenized" by expanding the scope of the knowledge/modeling? This is also the distinction between black swans and DK.

Mechanisms for dragon-kings

DKs may be associated with the regime changes, bifurcations, and tipping points of complex out-of-equilibrium systems [259]. For instance, a catastrophe (fold bifurcation) of the global ecology [26] – where incremental loading has little impact, but beyond a threshold results in a dramatic change that is difficult to reverse – could be considered to be a DK that has many precursors as the system approaches the catastrophe. Secondly, positive feedback, e.g,. where in a stampede the number of cattle running increases the level of panic which causes more cattle to run, can cause DK in crowds and stock markets. Next, attractor bubbling, as shown in Fig. 4.2, is a generic behavior appearing in networks of coupled oscillators whereby noise occasionally pushes the system trajectory into extreme orbits [45]. A block and spring mechanical model, considered as a model of geological faults and their earthquake dynamics, produces DK in a similar fashion [229]. It could also be the case that DKs are created as a result of system control or intervention. That is, trying to suppress the release of stress or death in dynamic complex systems may lead to an accumulation of stress or a maturation towards instability. For instance, brush/forest fires may be unnaturally suppressed, allowing for a buildup of

dead wood, and resulting in a huge uncontrollable fire. An analogue to this is monetary policy, in which quantitative easing programs and low interest rate policies aimed at smoothing out economic fluctuations lead to instability and an enormous DK "correction" event [249, 250].

Assessment, modeling, prediction & control

Prior to any discussion of prediction, it is important to mention that the identification and explanation of DK events is already difficult, and extremely important for risk assessment/management. As was mentioned, one can identify outliers in historical datasets, and attempt to associate them with a mechanism/process. This identification is important as it demonstrates the relatively high probability (or even possibility) of an event relative to a model, or the rest of the data. For such extreme events, data will be sparse, and uncertainty deep/severe [19]. However, if there is suggestive evidence of DK, then one should not simply ignore them in the name of parsimony. How one proceeds will depend on the focal application. However, one should generally be able to hypothesize relevant DK mechanisms and consider models for them, construct extreme scenarios (e.g., the cost of a severe meltdown at a nuclear power plant in Switzerland or New York state), try to deduce some probabilities for such events, assess uncertainties, and evaluate the sensitivity of the risk to these quantities. Methods for dealing with deep/severe uncertainty [19] can provide guidelines. To account for risk over future time periods one may consult approaches from risk theory whose compound process models consider both random event occurence and size [182].

Modeling DK requires dynamic models that are complex and/or non-linear, that are being fit to data provided by the continual monitoring of the focal system. For instance, in non-linear systems with phase transitions at a critical point, it is well known that a window of predictability occurs in the neighborhood of the critical point due to precursory signs: the system recovers more slowly from perturbations, autocorrelation changes, spatial coherence increases, etc.[269, 222]. These properties have been used for prediction in many applications ranging from changes in the bio-sphere[26] to rupture of pressure tanks on the Ariane rocket [16]. For the phenomena of unsustainable growth (e.g., of populations or stock prices), one can consider a growth model that features a finite time singularity, which is a critical point where the growth regime changes. In systems that are discrete scale invariant, such a model is power law growth, decorated with a log-periodic function [238, 128].

This has been applied to many problems [241, 16, 140] such as earthquakes [220], and the growth and burst of bubbles in financial markets [142, 262, 253, 86, 261]. Next, epidemic dynamics are interesting and may reveal the development of a block-buster success: e.g., the spread of plague, viral phenomena in media, the spread of panic and volatility in stock markets, etc. In such a case, a powerful approach is to decompose activity/fluctuations into exogeneous and endogeneous parts, and learn about the endogenous dynamics that may lead to high impact bursts in activity [61, 245, 92].

Regarding prediction, given the relevant model estimates, one may compute quantities such as the probability of an extreme (e.g., a DK) in a future time interval, related risk measures such as value at risk and expected shortfall [176], the most probable occurrence time of an event, etc. In a dynamic setting the dataset will grow over time, and the model estimate, and its estimated probabilities, will evolve. Tests for DK (or anomolous dynamics in general) will likely be weak most of the time (e.g., when the system is around equilibrium), but as one approaches a DK, and precursors become visible (i.e., as one enters a pocket of predictability), the true positive rate should increase.

Regarding control, one can then use the estimated probabilities and their associated uncertainties to inform decisions such as taking a specific action if a DK is predicted to occur. An optimal decision (e.g., see [31] for the statistical decision theory) will then balance the cost of false negatives/false positives and misses/false alarms according to a specified loss function. For instance, if the cost of a miss is very large relative to the cost of a false alarm, then the optimal decision will detect DKs more frequently than they occur. One should also study the true positive rate of the prediction. The smaller this value is, the weaker the test, and the closer one is to black swan territory. It has been proposed that the more homogenous and connected the system, the more predictable its behavior will be [258], as presented in Fig. 4.4. An important point to mention and ponder is the risk and potential harm resulting from well intended control. Recall that the suppression of natural fluctuations in a complex system is itself a source of DK events.

To mention some critical statistical issues: Second only to the selection of the proper model is selection of the relevant variables. For this one should consider scientific hypotheses as well as modern statistical/machine learning methods – in particular when predictive power is emphasized over the inferential value of a precise and simple model. Next, in any such modeling (of extremes with non-trivial models) there is bound to be substantial uncertainty. Thus, one should adequately capture not only the randomness present in the fitted stochastic model, but also the uncertainty of

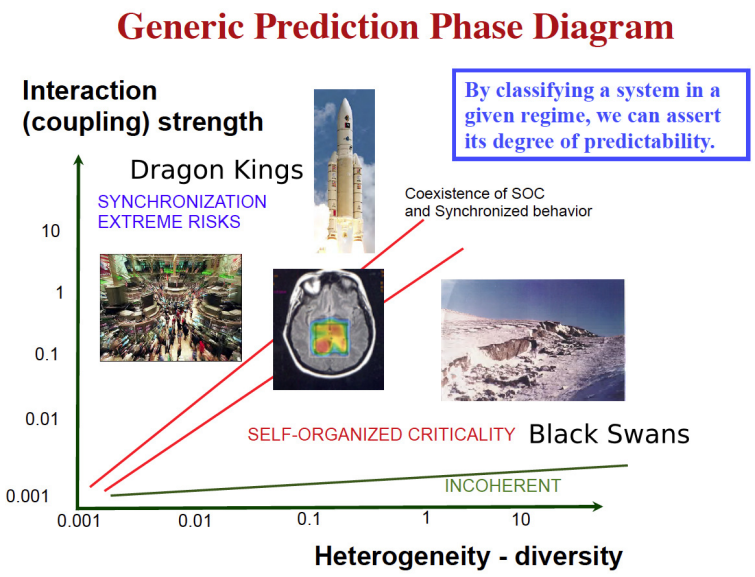

Figure 4.4: Predictability based on interaction and diversity in a system. This illustration has been reproduced from [258].

its estimated parameters (e.g., with Bayesian techniques), and the uncertainty in model selection (e.g., by considering scenarios / an ensemble of different models). Thirdly, in the absense of data, when considering scenarios, to obtain the sampling distribution of model parameters, or to optimize prediction/decision algorithms, simulation is a powerful tool effectively allowing one to conduct rigorous quantitative thought experiments, and providing a valuable complement to the analysis.

The impact of extremes, and DK in particular, urges that extremes be studied and monitored. Future designs should be robust and resilient [255], acknowledging the potential for such extremes. At the same time, the application of such theory presents many challenges and demands modesty in the presense of severe uncertainty.

Chapter 5

A general test for multiple outliers

This chapter is based on [290].

5.1 Introduction

There is a large statistical literature on outlier testing/detection (e.g., [25, 120] are classic references). Outliers are anomolous observations that may be a spurious nuissance to be discarded, or may be of primary relevance: e.g., in climate science, credit card fraud detection, medical diagnosis, etc. (see [10] for examples). Much of the statistical literature focuses on testing outliers relative to a *null* model (i.e., a model without outliers) that is Normal. In many cases, the outliers are themselves considered to be Gaussian. For instance, a Gaussian sample with Gaussian outliers is the so-called *contaminated (location-shift) normal*, which is the "the simplest, and perhaps most studied, case" [280], classically employed to analyze the performance of outlier tests in Gaussian samples (Sec. 3.4 of [120]). Furthermore, a mixture of Gaussian densities has often been used for the detection of outliers/anomoulous sub-populations [125, 12]. But most empirical data of interest in many fields do not follow Gaussian distributions but are better described by distributions with fatter tails such as exponential or power law distributions [158, 183, 189, 243].

Here, we thus consider the detection of outliers in samples having approximately exponential or Pareto tails. The case of an exponential null DF (distribution function) has been covered in literature, and [23] provides a review. This case is much more general than typically claimed. In particular, by a simple transformation, the outlier tests in an exponential sample are applicable to Pareto (power law) samples. Further, Extreme Value Theory (EVT) [79] provides that general "well behaved" DFs

asymptotically have either exponential or Pareto tails. Thus, testing outliers relative to exponential tails is very general because EVT provides asymptotic generality for the null DF. Finally, the tests considered here are independent of the parameter of the exponential/null DF. If a different DF were considered, this would require one to estimate the parameters of the DF in the presence of outliers to perform the test, which is bound to lead to strong biases.

A common approach to outlier detection, with which the optimality of outlier tests in exponential samples is studied, is the so-called *slippage model*, where the outliers also have an exponential DF, but with a larger scale parameter than the null [120, 25]. However, there is no reason why the DF of the outliers needs to be a scale multiple of the null DF, aside from the fact that this simplifies analysis. Rather many shapes should be possible, reflecting the fact that outliers arise not only from amplification dynamics (of which there are many), but also human error, instrument error, or any other arbitrary departures from ordinary system behaviour. Therefore, the performance of various tests with different outlier generating mechanisms should be compared to obtain the most general outlier detection methodology.

For a statistical outlier test, one not only wants to have high power and computational tractability, but also to estimate the number k of outliers well. *Masking* and *swamping* errors are impediments to this task: **Masking**: For k actual outliers, and r hypothesized outliers, with $r < k$, a first outlier *masks* a second if the second outlier is only identified as an outlier when the first is not present. That is, considering $r < k$ outliers, $k - r$ outliers have been left in the sample, and may skew the statistics enough so that the r hypothesized outliers do not appear very extreme. The larger the $k - r$ outliers remaining in the sample, the worse the masking.

To quote [120] on masking: "*the masking effect is the main cause, both of the large degree of attention given in the literature to the detection of multiple outliers, and the fact that none of the solutions proposed is entirely satisfactory.*"

Swamping: For k actual outliers, and r hypothesized outliers, with $r > k$, an outlier *swamps* a non-outlier when the non-outlier is only identified as an outlier when considered in the presense of the outlier. That is, when $r - k$ non-outliers are grouped with the k outliers within a TS (test statistic(s)), the test may still reject the null hypothesis, especially if the k outliers are large.

That block outlier tests suffer from both masking and swamping has led to the introduction of sequential testing [215, 120, 147, 25]. Inward (sequential) tests were found to suffer from masking, and

thus outward sequential tests were proposed [215, 147] and have since become the standard approach. However, they are substantially more complicated as they require a multiple testing correction to control the type 1 error. In this work, it is shown that a simple modification to the TS cures the inward test of the masking problem, making it competitive with the outward test.

There are many works in the literature where a limited set of TS and testing methods are compared with specific null and alternative models (e.g., [169, 50, 24, 170]). Here, a more comprehensive comparison is done for a range of these statistics and of null and alternative models, providing useful practical insights that – in the opinion of the authors – have not been emphasized in the literature.

Finally, the present article aims at shifting the focus from reliability/failure applications (the exponential case) towards applications in risk modeling (the Pareto case). Indeed, Pareto (power law) DFs seem to be ubiquitous in most natural hazards (earthquakes, landslides, floods, tsunamis, etc.), industrial catastrophes (nuclear accidents, hydro-electric dam ruptures, power black-outs, traffic grid-locks, etc.), social systems (individual wealth, size/success of companies, etc.), and so on (see e.g. [183, 189, 243] and references therein). Furthermore, a number of studies have found suggestive evidence that there are extreme events "beyond" the Pareto sample [246, 259]. This brings into play the concept of "Dragon Kings" (DK) [246], which will be elaborated. We should stress again that, via a simple transformation, the outlier tests in an exponential sample can be directly translated to Pareto (power law) samples and vice-versa.

Section 5.2 presents the general methodology, its justification, and a battery of statistical tests for the detection of outliers. This includes a modification of the classical TS, as well as the presentation of general arguments based on EVT that supports the generality of the exponential outlier test. In Section 5.3, a variety of Monte Carlo studies are done to compare the performance of the different tests, taking into consideration both dispersed and clustered outliers, susceptability to masking and swamping, and robustness to null misspecification. In Section 5.4, a case study is given, and the Dragon King (DK) concept is explained. In the supplementary material four additional case studies are presented, which highlight results from previous sections. The case studies are: financial crashes (drawdowns), nuclear power generation accidents, stock returns, fatalities in epidemics, and city sizes. Section 5.5 concludes with a discussion.

5.2 Outlier Testing Methodology

The setup is an ordered sample $x_{(1)} > x_{(2)} > ... > x_{(n)}$ where $n - k$ of the observations are iid (independent and identically distributed) realizations of a random variable, $X \overset{\text{iid}}{\sim} \text{Exp}(\alpha)$, with the exponential DF,

$$F_X(x) = 1 - \exp\{-\alpha x\},\ x \geq 0,\ \alpha > 0\ , \tag{5.1}$$

and the remaining k points are outliers that are iid with some DF, and independent of X. Which points are outliers is unknown, and one wants to detect them. Further, if X is exponentially distributed then $Y = u\exp\{X\} \overset{\text{iid}}{\sim} \text{Pareto}(\alpha, u)$. That is, the exponential of an exponential random variable has the Pareto DF,

$$F(x) = 1 - (x/u)^{-\alpha},\ x \geq u,\ \alpha > 0\ . \tag{5.2}$$

Therefore one can take the logarithm of Pareto samples and apply outlier tests for exponential samples. The following subsections discuss block and sequential testing procedures, different TS, the justification for why exponential tails are general, and how to identify at what threshold the exponential tail begins.

5.2.1 Block, inward, & outward tests

A simple approach to detecting outliers is a *block* test, where the number of outliers, r, is specified a-priori and, in a single test, either all r points in the block are identified as outliers or zero are. Such procedures suffer from masking and swamping when too many or too few points are included in the block respectively. However, if well specified, block tests are powerful due to the simultaneous usage of all data. To avoid dependence on the specification of block size r, sequential tests were developed:

Inward test: One starts with the full sample and tests if the largest point is outlying. If that point is identified as outlying (the test is rejected), then the point is removed from the sample and the test is repeated with the next largest point. The procedure is repeated until the first failure to reject. The estimated number of outliers $\widehat{k}$ is the number of rejected (marginal) tests. Clearly, this test can suffer from both masking and swamping. The weaknesses of the inward procedure were cited as motivation for the *outward* test [215, 120, 147]:

Outward test: One specifies a maximum number of outliers r, and starts by testing if the rth largest point $x_{(r)}$ is an outlier by deleting the other $r - 1$ largest values $x_{(r-1)}, x_{(r-2)}, ..., x_{(2)}, x_{(1)}$ and applying the test on $x_{(r)}$. If this test is rejected, then r outliers are identified. If this test is not

rejected, then one takes a step "outward", which involves then testing the $(r-1)$th largest point $x_{(r-1)}$. This testing of increasingly large points is done until the first rejected test, say for $x_{(j)},\ j \in \{1, ..., r\}$, thus identifying $\widehat{k} = j$ outliers. If none of the tests are rejected, then no outliers are identified. This test minimizes the probability and magnitude of both masking and swamping. As such, the outward procedure has been claimed superior over the inward [147, 50, 24] and received more subsequent development [169, 170].

However, control of the type 1 error (the probability of a false alarm) is difficult in the outward test. The test considers the null hypothesis H_0 that there are no outliers, with multiple alternatives, H_j that there are j outliers $j = 1, ..., r$, with TS T_j. A single rejection of the r tests rejects the null H_0. Thus, to achieve an overall type 1 error level of $0 \leq a \leq 1$, e.g., the common level of 0.05 or 0.1, the marginal tests need to have a lower level. The larger r is, the larger the correction will be, and thus the lower the power of the test. This "multiple testing correction" requires knowing the joint and marginal DF of, generally dependent, $T_j,\ j = 1, ..., r$. More specifically, one defines all marginal tests to have equal level b, i.e., $Pr\{T_j > t_j\} = b,\ j = 1, ..., r$, and the level b is determined such that $Pr\{T_j \leq t_j,\ j = 1, ..., r|H_0\} = 1 - a$. Clearly $a^r \leq b \leq a$, where the lower bound corresponds to the case of independent tests (the Bonferonni bound), and the upper bound to perfect dependence. For the specific TS (5.3) discussed below, the joint and marginal DFs were derived for $k = 2, 3$ in [147], and a Monte-Carlo implementation recommended in [169] for larger k.

In contrast, for the inward method, the type 1 error level is equal to the marginal level ($a = b$) because a rejection of the null only happens when the first marginal test (for the largest point, $x_{(1)}$) is rejected. This is a major advantage over the outward procedure in terms of computation and also because no power is lost due to a multiple testing correction.

5.2.2 Gallery of test statistics

We now review the standard TS for outlier detection in exponential samples, and propose a modification to cure inward tests of the masking error. In general, outlier TS facilitate a comparison of the "outlyingness" of the suspected outliers (in the numerator of the statistic) relative to some measure of dispersion within another subset of the data (in the denominator of the statistic). Some of the measures are based on spacings (or maxima) and others on sums of observation sizes.

The *max-robust-sum* (MRS) statistic for the j^th rank,

$$T_{j,m}^{MRS} = \frac{x_{(j)}}{\sum_{i=m+1}^{n} x_{(i)}}, \quad m \geq 0 \ , \tag{5.3}$$

is a modification of a classic statistic [147], which is recovered when $m = 0$, where m is a pre-specified maximal number of outliers. When $m = 0$ it is referred to as the MS statistic. The index j is given to allow the test to be used in sequential procedures, for $j = 1, ..., m$ [147]. The case of $m > 0$ in the denominator was introduced to prevent masking: When $r < k$, and $m = 0$, there will be $k - r$ outliers in the denominator that will make $x_{(j)}$ appear less outlying. This becomes important in inward testing, and is similar to using robust scale estimates in the case of outliers relative to a Normal population [136]. Thus, the choice of m is a tradeoff between sample size (power) and sample purity (masking avoidance). The classic statistic (when $m = 0$) has optimal properties in the presense of a single outlier from an exponential DF [120]. Having a single value in the numerator rather than a sum (as in eq. 5.4), this statistic will not cause swamping. That is, $x_{(j-1)}$ being outlying has no influence on the test for its smaller neighbour $x_{(j)}$. However, the downside of this, that will be seen, is that this statistic is not powerful when outliers are clustered. The DF of the test statistic under the null is conveniently computed by Monte Carlo simulation. The test statistic (5.3) may be referred to as the MS test when $m = 0$.

We propose the *sum-robust-sum* (SRS) test statistic for r upper outliers,

$$T_{r,m}^{SRS} = \frac{\sum_{i=1}^{r} x_{(i)}}{\sum_{i=m+1}^{n} x_{(i)}}, \quad m \geq 1 \ , \tag{5.4}$$

which is a modification of the well known *Cochrane* test statistic, which is recovered when $m = 0$, where m is again a pre-specified maximal number of outliers. When $m = 0$, it may be referred to as the SS statistic. In the classical case ($m = 0$), the test is equivalent to a likelihood ratio test (LRT) when the outliers also come from an exponential [23]. Due to the sum over r in the numerator, this test suffers from swamping. However it is not susceptible to masking because it uses the observation magnitude rather than differences; i.e., it does not compare $x_{(1)}$ versus $x_{(2)}$, which may be close to each other, but far from the rest of the sample. Further, by summing in the numerator, it will also be powerful in the detection of cases where the outliers are clustered. The DF of this statistic (when $m = 0$) was given by [167, 50]. It is also convenient to simulate the DF. The test statistic (5.4) may

be referred to as the SS test when $m = 0$.

Another classic test statistic, for r upper outliers, is the *Dixon* (D) statistic [73],

$$T_r^D = \frac{x_{(1)}}{x_{(r+1)}} \ , \tag{5.5}$$

whose DF under the null is given by [168]. In the outward testing case, the joint DF was given by [170]. It is often used as a less powerful alternative to the SS, with the advantage of being less prone to both swamping and masking.

We also include a test from the Physics literature on detecting "Dragon King" (DK) outliers [201]. This *DK* statistic for r upper outliers,

$$T_r^{DK} = \frac{\sum_{i=1}^{r} z_i}{\sum_{i=r+1}^{n} z_i} \ \sim F_{2r,2(n-r)}, \tag{5.6}$$

uses the weighted spacings, $z_i = i(x_{(i)} - x_{(i+1)}),\ i = 1, ..., n-1,\ z_n = nx_{(n)}$, and has an F distribution. It suffers from both masking and swamping, and will not be powerful in the case of multiple clustered outliers since it counts spacings rather than absolutes. It is thus only superior in the simplicity of its DF under the null.

Under the exponential DF (5.1), all of these TS have the pleasant property that their DF is independent of α. Under a different DF, the DF of the TS would depend on the parameters of the null that would then need to be estimated, potentially in the presence of outliers. This follows from the Rényi representation of spacings [211, 23] where for $E_i \overset{\text{iid}}{\sim} \text{Exp}(\alpha)$, the spacings $S_i = X_{(i)} - X_{(i-1)}$ are equal in distribution to $(\alpha i)^{-1} E_i$ where $E_i \overset{\text{iid}}{\sim} \text{Exp}(1)$. Thus, in the TS, which are ratios of sums of spacings or order statistics (which are themselves a sum of spacings), the parameter α cancels. In this work, with the exception of the DK test (5.6), the empirical DF of the TS are computed from 50,000 independent samples from the null DF.

In addition to the tests mentioned above, a mixture model is considered,

$$f(x) = (1-\pi)\alpha \exp\{-\alpha x\} + \pi\phi(x; \mu, \sigma) \ , \ \alpha, \sigma > 0 \ , \tag{5.7}$$

where the Gaussian density $\phi(x; \mu, \sigma)$ provides the outlier regime, and $0 \leq \pi \leq 1$ is a weight. It is common and natural to consider Gaussian outliers [280, 12, 125]. This model allows one to classify

points as either outliers or not based on the relative mass of the mixture components at that point. The MLE (Maximum Likelihood estimate) of this model (5.7) is done using an EM (Expectation Maximization) algorithm [210]. A LRT (likelihood ratio test) of this model versus the null ($\pi = 0$) provides p-values, and $n\hat{\pi}$ estimates the number of outliers. The major strengths of this method are that it does not require sequential testing – i.e., it avoids masking and swamping naturally – and that one can generalize the exponential, e.g., to a Weibull or gamma DF, without complicating the procedure. It is important to note that this method does not distinguish between inliers and outliers – i.e., the density ϕ can be significant both within and beyond where the null df has substantial mass.

5.2.3 EVT & outlier testing in exponential tails

It is important to note that outlier tests based on both the Pareto and exponential DFs are generally applicable to data having approximately Pareto or exponential tails. This follows from the well known Pickands-Balkema-de Haan theorem of EVT, that states [79]: For a broad range of DFs, for random variable X, with sufficiently high threshold u, the excess DF, $F_u(x) = P\{X - u \leq x | X - u > 0\}$ (i.e., the tail of the DF), is approximated by the GPD (Generalized Pareto Distribution Function),

$$GPD(x; \xi, \beta, \mu) = \begin{cases} 1 - (1 - \xi(x-\mu)/\beta)^{-1/\xi}, & \text{if } \xi \neq 0 \\ 1 - \exp(-(x-\mu)/\beta), & \text{if } \xi = 0 \ , \end{cases} \tag{5.8}$$

in the sense that,

$$\lim_{u\to\infty} \sup_{0\leq x} | \ F_u(x) - GPD\left(x | \xi, \beta(u), \mu\right)| = 0 \ , \quad \beta(u) > 0, \forall u. \tag{5.9}$$

If $\xi = 0$ (the Gumbel case), then the GPD (5.8) is exponential with lower truncation $\mu = u$ and scale parameter $\beta = 1/\alpha$. This case includes common DFs such as the exponential (obviously), the Normal, and even some fat-tailed DFs such as the Lognormal. If $\xi > 0$ (the Fréchet case), the GPD (5.8) is (generalized) Pareto with $\mu = u$, $\sigma = u/\alpha$, and $\xi = 1/\alpha$. This case includes heavy tailed DFs such as the Pareto and Log-gamma. The only other case ($\xi < 0$: the Weibull case) is where the DF has a finite upper endpoint, which is of less interest in outlier detection. Therefore, since a Pareto tail can be transformed to an exponential one, outlier testing in exponential samples is (asymptotically) extremely general!

Since the GPD approximation (5.9) is only asymptotically valid, one must select a sufficiently large lower threshold u before applying outlier tests. The problem of threshold selection is a tradeoff between bias and variance, and is the primary statistical issue in the EVT literature, where it is referred to as sample fraction selection. In the physics literature, threshold selection and goodness of fit diagnostics are important for the interpretation of mechanisms underlying power laws found in datasets. There are a variety of tools available for this task.

The classic "Hill plot" method [122] for threshold selection consists of estimating the model for a range of thresholds and selecting the lowest threshold (the largest sample fraction) where the estimate is "stable" – i.e., consistent with values of the estimate for larger thresholds. See Fig. 5.5 for an example. Of course, one can also look for statistically significant changes in the estimated parameter relative to the hopefully stable value obtained deeper in the tail [122, 109, 28], however more powerful principled methods exist (see e.g., [30, 105] for a review). For instance, let us mention the methods based on minimizing the asymptotic mean square error of the estimate. This requires assuming the (class of) DF beyond the power law tail [109], or using bootstrap methods [63, 105].

For some reason, these methods have not been extensively adopted outside of the EVT literature. For instance, the most highly cited paper on the estimation of power laws and sample fractions [52] does not mention the sample fraction estimation literature. However a subsequent work [283], extending the method to binned/aggregated data, does provide such references. The popular work [52] suggests choosing the pair of u and α that have the smallest KSD (Kolmogorov-Smirnov distance). The KSD criterion penalizes error, and rewards sample size. However, as noted by [70, 56], comparing KSD across samples of different size is not necessarily consistent as the KSD simply scales with growing sample size like $\sim 1/n^{0.5}$. Further, in [52], no argument was given why this is optimal. In [70, 56], it was shown that the method fails when the DF has a power tail whose parameter changes from one value to another. Originally, [122, 109] proposed applying a test for decreasing u, and selecting u at the value before the first value where the test is rejected. In [70], a similar approach was proposed based on the KS test, where instead one would select the largest sample that could not be rejected, regardless of if rejection occurs at higher thresholds.

These methods can be thought of as outlier tests, where "lower outliers" are points below the tail threshold u, that are discordant with the tail. However, instead of elaborating on this, a more general automatic approach is recommended: One should fit both the exponential, and a more complicated

density to the range of upper samples, and identify the threshold at which the complicated density is not significantly better. If the more complicated density is sufficiently flexible, this should determine that the exponential provides a good approximation above the threshold. One could consider comparing nested models with the LRT, however this is only a comparison with a specific alternative model. For a more general alternative model, one can use a non-parametric estimator, such as the logspline estimator (available in R:locfit) [154]. One can then compare the null with this alternative with the Akaike Information Criterion (AIC). In the presense of clear outliers, one may wish to use estimators that censor, or are robust to the outliers.

Concerning outlier testing, it is useful to estimate the sample fraction to have an idea of where the tail approximation begins to apply. However, tests can often accept a model for a larger sample but reject it in the tail! Thus, one should apply outlier tests for a range of lower thresholds and look for stability in outlier test results for data that do not violate the null. That is, letting $n_u \leq n$ be the size of the largest upper sample that can be defended based on the methods discussed above, an outlier test should be applied to the upper samples consisting of the $n_u, n_{u-1}, ..., 10+r$ largest points, where r is the expected number of outliers, and where one should certainly not consider samples of size smaller than ten. Consistent identification of outliers in these upper subsamples, where the GPD approximation (5.9) is most relevant, and where the null model cannot be rejected, should be interpreted as a robust result. This algorithm involves $c = n_u - (10 + r)$ consecutive dependent tests, which gives multiple chances for a false positive. However, under the null, the probability of rejecting $c > 1$ consecutive tests, decreases as c increases. Based on simulation studies with the range of models considered within this work, we offer as a rough rule of thumb, that for a sample of size $10 < n < 100$, one should require a run of $c = n/10$ tests to be rejected to maintain control of the type 1 error.

5.3 Outlier Test Performance

Here, the performance of the different tests are compared with simulation studies. The strength of TS are studied in block tests and compared with the mixture test; where masking and swamping are studied in block tests; where inward, outward, and mixture tests are compared; and how misspecification of the null effects test performance. The setup considered is a standard exponential sample with four outlier scenarios: (0) no outliers, (I) a single outlier, (II) multiple dispersed outliers, and (III) a cluster of multiple outliers. These cases are plotted in Fig. 5.1.

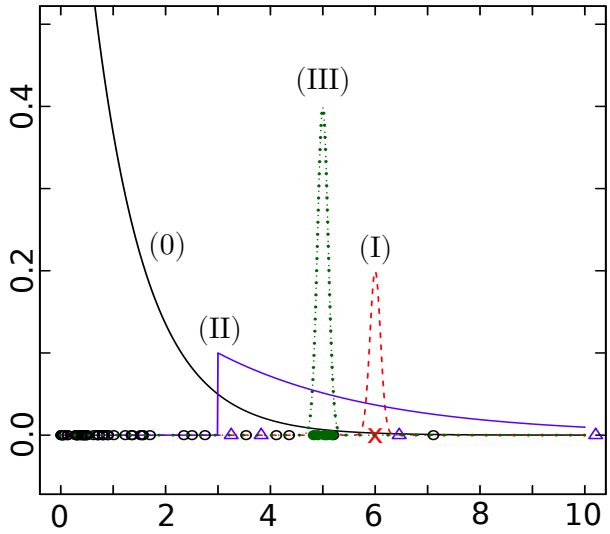

Figure 5.1: **Outlier cases**. The null case (0) is standard exponential for which a realization of 50 points are plotted as open circles. Three outlier cases are considered on top of the null: (I) a normal DF with mean $\mu = 6$ and $\sigma = 0.1$ is given by a dashed red line, and its single outlier is the red x mark; (II) multiple dispersed outliers $Y_i \sim 3+\text{Exp}(1/\beta),\ i = 1, ..., 5$ plotted with a solid blue line for $\beta = 4$ and blue triangles indicating (a realization of) the outliers; (III) multiple clustered outliers $Y_i \sim \text{Norm}(\mu, 0.1),\ i = 1, ..., 5$ plotted with a green dotted line for $\mu = 5$, and green dots indicating the outliers.

5.3.1 Block test performance compared with a mixture model

Here, the power (at level 0.1) of the range of TS is studied where the TS are employed in block tests, and where the block size r, and the robustness value m are set to the true number of outliers k. The three cases where outliers are present are considered: (I) $n = 20$, $k = 1$, $X_i \sim \text{Exp}(1)$, $i = 1, ..., 19$, $X_{20} \sim \text{Norm}(\mu, 0.1)$, $\mu = 3, ..., 10$; (II) $n = 50$, $k = 5$, $X_i \sim 3+\text{Exp}(1/\beta),\ i = 46, ..., 50$, $\beta = 1, 2, ..., 6$; (III) $n = 50$, $k = 5$, $X_i \sim \text{Exp}(1)$, $i = 1, 2, ..., 45$, $X_i \sim \text{Norm}(\mu, 0.1),\ i = 46, ..., 50$, $\mu = 3, 4, ..., 10$. The

mixture model (5.7) is only estimated in the cases with multiple outliers.

In Fig. 5.2, the power curves are plotted for these scenarios, for a range of outlier parameters, being computed over 2'000 independent simulations. For a single outlier (case I), most of the tests are exactly identical (by definition), with the exception of the DK and D tests, which are weaker. For multiple dispersed outliers (case II), the SS test performs best, with the SRS being slightly less powerful – robustness has a cost. The mixture is poorly specified and is thus weakest. For clustered outliers (case III), the performance of the tests varies greatly. Indeed, the TS with the sum in the numerator often identifies the cluster of outliers. However, the well specified mixture model is most powerful, also identifying the "outliers" when they are not really outlying but rather a contamination well within the sample (i.e., "inliers").

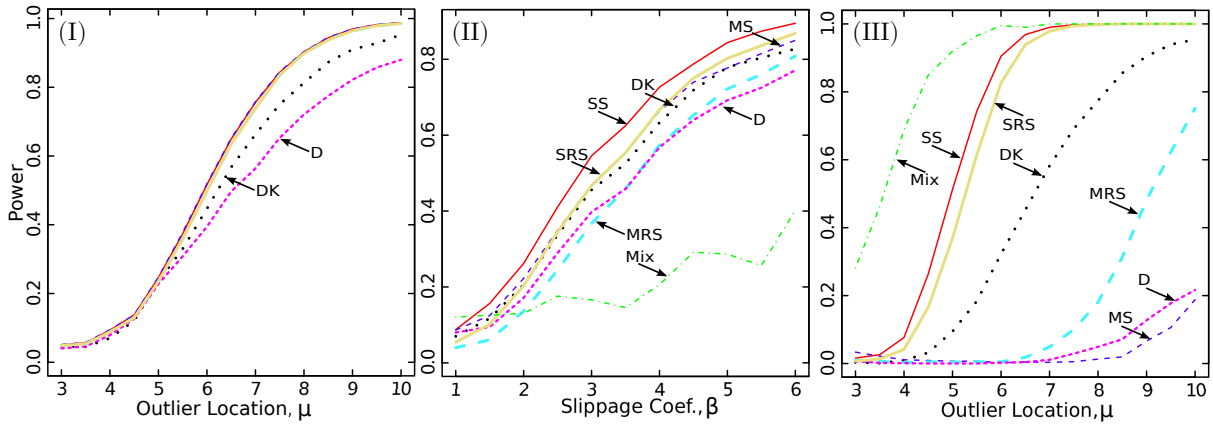

Figure 5.2: The Monte Carlo power curves (at level 0.1) for the range of test statistics (labeled) employed in block tests, providing the rejection rate for different values of the outlier parameter. Each panel corresponds to one of the outlier cases defined in figure 5.1.

5.3.2 Masking, swamping, and estimating the number of outliers

We now present simulation studies to expose the degree to which the different TS suffer from masking and swamping in block tests – that is, how accurately they estimate the number of outliers. This is done by performing the tests on synthetic data for a range of block sizes. The three scenarios considered are: (I) swamping due to a single outlier, $n = 30$, $k = 1$, $X_i \sim$Exp(1), $i = 1, ..., 29$, $X_{30} \sim$Norm$(8, 0.1)$; (II) swamping without masking due to dispersed outliers, $n = 30$, $k = 5$, $X_i \sim$Exp(1), $i = 1, ..., 25$, $X_i \sim$ 3+Exp(1/5), $i = 26, ..., 30$; and (III) swamping with masking due to clustered outliers, $n = 30$, $k = 5$, $X_i \sim$Exp(1), $i = 1, ..., 25$, $X_i \sim$Norm$(8, 0.1)$, $i = 26, ..., 30$.

Our simulation study determines the frequency at which the tests are rejected, at level 0.1, in 2'000 independent samples, for a range of block sizes ($b = 1, 2, ..., 10$). The results are in Fig. 5.3. The MS and MRS tests are not effected by block size since the maximum is always the largest point. In the next section, the inward test will apply the MRS statistic to the largest point, then the second largest, and so on. In that case, the MRS TS will not cause swamping. As anticipated, masking is problematic for the MS statistic, especially when large observations are densely clustered. Further, as intended, the MRS suffers from masking less than the MS. The SS and SRS tests suffer less from masking and swamping than those based on spacings and maxima. Swamping is pervasive in block testing, even when there is only a single large outlier. That the rejection rate decays slowly as the block size surpasses the true block size indicates that the minimal p-value in the sequence of estimates will not reliably indicate the true block size. These problems motivate sequential testing.

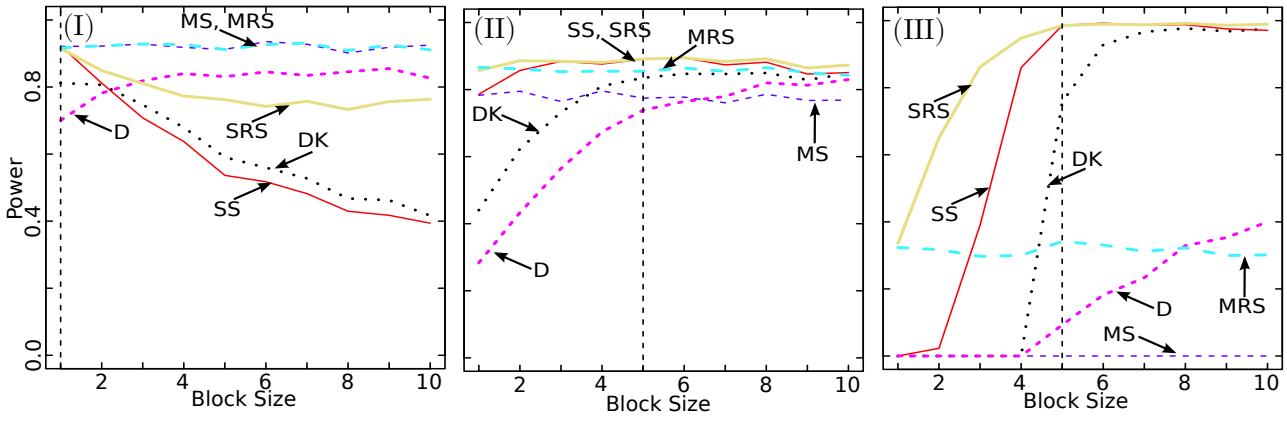

Figure 5.3: Monte carlo rejection rate (at level 0.1) for the range of test statistics (labeled) employed in block tests, with different block sizes. Each panel corresponds to one of the three outlier cases.

5.3.3 Comparative study of the performance of sequential estimators

Here, inward and outward sequential procedures are compared, along with the mixture test. Again the four outlier scenarios visualized in Fig. 5.1 are considered. The tests used are: (i) the outward test with MS, MRS, SS, and SRS statistics; (ii) the inward test with only the MRS statistic, which is necessary to avoid masking and swamping; (iii) the mixture model (5.7); and (iv) the SRS block test, given the correct number of outliers. This last option, which was the best performing block test in Fig. 5.2, provides a benchmark.

The DFs for the TS were simulated with 50'000 samples from the null model. All tests were

done with a level of 0.1. For the outward test, the level of the marginal tests b was lowered to obtain the overall level of $a = 0.1$. For each test, this was done by applying the test on 10'000 independent samples generated from the null, for multiple values of b, and selecting b such that $a(b) = 0.1 \pm 0.005$. The resultant marginal levels are in Table 5.1. Note how large of an adjusment is needed in the outward test, whereas in the inward test there is no adjustment: $b^{\text{Inward}} = a = 0.1$.

n	r	MS	SS	MRS	SRS
50	10	0.018	0.05	0.025	0.049
30	5	0.028	0.055	0.0345	0.0575
15	5	0.025	0.06	0.036	0.056

Table 5.1: Marginal levels (b) for outward tests for different sample sizes (n), maximal number of outliers (r), and robustness value ($m = r$) to obtain an overall type 1 error level of $a = 0.1$

The results, for slightly different specifications of the four cases, and in order of decreasing sample size, are in Tables 5.2, 5.3, and 5.4. In case (0), where there are no outliers, the inward and mixture procedures have false positive events that estimate a small number of outliers, whereas the outward procedures falsely identify large numbers of outliers. In case (I) of the sequential procedures, the inward test is most powerful at identifying the single outlier, even matching the power of the block test. The outward tests are substantially weakened, even with relatively small $m = 5$. The inward test provides superior estimation of outliers, whereas the other tests tend to overestimate. In case (II), with a cluster of outliers, both the benchmark (the block test) and the inward test perform poorly. They are outperformed by the outward test, which is less susceptible to masking, by design. However, here the mixture approach is both the most powerful and accurate in estimating outlier numeracy. In case (III), with multiple dispersed outliers, all of the inward and outward approaches are similarily competitive, while being slightly dominated by the block test. The mixture approach is weak since the outlier component is poorly specified. For the outward procedure, the MS/MRS statistic dominates the SS statistic.

In summary, the inward procedure with the MRS test statistic is more computationally convenient than the outward procedure, commits less severe false positives, and can even be more powerful when identifying single or multiple dispersed outliers. In the event of a dense cluster of outliers, a mixture approach can be more computationally convenient and powerful than the outward approach. Within the outward approach, the MS/MRS statistic was superior to the SS/SRS statistic, and robust modifications performed similarly.

Case	Quantity	MS Out	SS Out	MRS Out	SRS Out	MRS In	Mix	SRS Block
(0)	Rej. Rate	0.11	0.10	0.11	0.10	0.10	0.14	0.10
(0)	$\widehat{k}$	(3,6,9)	(5,9,10)	(3,6,9)	(5,9,10)	(1,1,3)	(2,2,4)	
(I)	Rej. Rate	0.30	0.22	0.30	0.22	0.64	0.09	0.69
(I)	$\widehat{k}$	(2,3,6)	(2,5,10)	(2,3,7)	(2,5,10)	(1,1,2)	(2,2,2)	$= 1$
(II)	Rej. Rate	0.91	0.75	0.89	0.75	0.04	0.95	0.38
(II)	$\widehat{k}$	(5,7,8)	(5,7,10)	(5,7,9)	(5,7,10)	(1,9,10)	(5,5,6)	$= 5$
(III)	Rej. Rate	0.96	0.96	0.97	0.96	0.95	0.63	0.98
(III)	$\widehat{k}$	(5,6,8)	(4,6,10)	(5,6,9)	(4,6,10)	(6,7,10)	(3,10,10)	$= 5$

Table 5.2: $n = 50$ (sample size), $m = 10$ (robustness value). Summary of tests over 5000 repeated simulations of four cases: (0) the null case ($X \sim$Exp(1)), (I) a single large outlier ($X_i \sim$Exp(1), $i = 1, ..., 49$; $X_{50} \sim$Norm$(7, 0.1)$), (II) a cluster of multiple outliers ($X_i \sim$Exp(1), $i = 1, ..., 45$; $X_i \sim$Norm$(5, 0.1)$, $i = 46, ..., 50$); (III) multiple dispersed outliers ($X_i \sim$Exp(1), $i = 1, ..., 45$; $X_i \sim \max(\{X_i : i = 1, ..., 45\})+$Exp(1/5), $i = 46, ..., 50$). The rejection rate and the median $\widehat{k}$ and quartiles of the estimated number of outliers (in the event of a rejection) are given in alternating rows.

Case	Quantity	MS Out	SS Out	MRS Out	SRS Out	MRS In	Mix	SRS Block
(0)	Rej. Rate	0.11	0.11	0.11	0.11	0.11	0.16	0.10
(0)	$\widehat{k}$	(2,3,5)	(4,5,5)	(2,4,5)	(3,5,5)	(1,1,3)	(2,2,5)	
(I)	Rej. Rate	0.45	0.32	0.43	0.33	0.72	0.08	0.75
(I)	$\widehat{k}$	(1,2,3)	(1,3,5)	(1,2,3)	(1,2,5)	(1,1,2)	(2,2,2)	$= 1$
(II)	Rej. Rate	0.72	0.63	0.73	0.64	0.08	0.96	0.36
(II)	$\widehat{k}$	(3,4,5)	(3,4,5)	(3,4,5)	(3,4,5)	(4,5,5)	(3,3,3)	$= 3$
(III)	Rej. Rate	0.87	0.86	0.89	0.86	0.88	0.50	0.90
(III)	$\widehat{k}$	(2,4,4)	(2,4,5)	(2,4,5)	(3,4,5)	(3,4,5)	(2,5,7)	$= 3$

Table 5.3: $n = 30$ (sample size), $m = 5$ (robustness value). Summary of tests over 5000 repeated simulations of four cases: (0) the null case ($X_i \sim$Exp(1)), (I) a single large outlier ($X_i \sim$Exp(1), $i = 1, ..., 29$; $X_{30} \sim$Norm$(7, 0.1)$), (II) a cluster of multiple outliers ($X_i \sim$Exp(1), $i = 1, ..., 27$; $X_i \sim$Norm$(5, 0.1)$, $i = 28, 29, 30$), (III) multiple dispersed outliers ($X_i \sim$Exp(1), $i = 1, ..., 27$; $X_i \sim \max(\{X_i : i = 1, ..., 27\})+$Exp(1/5), $i = 28, 29, 30$). The rejection rate and the median $\widehat{k}$ and quartiles of the estimated number of outliers (in the event of a rejection) are given in alternating rows.

5.3.4 Robustness to null mis-specification

In practice, the correct specification of the null/main model is of considerable importance. Here, the sensitivity of the rate of false positives (level / type 1 error), and true positives (power), to the degree of misspecification of the null are exposed via a simulation study, for the battery of TS implemented in block tests. We consider simulating data from a Weibull DF,

$$F(x) = 1 - \exp\{-(x/\tau)^{\kappa}\},\ x \geq 0,\ \tau, \kappa > 0\ , \tag{5.10}$$

Case	Quantity	MS Out	SS Out	MRS Out	SRS Out	MRS In	Mix	SRS Block
(0)	Rej. Rate	0.11	0.11	0.11	0.11	0.08	0.16	0.10
(0)	$\widehat{k}$	(2,3,4)	(3,5,5)	(2,3,5)	(3,5,5)	(1,2,4)	(2,3,5)	
(I)	Rej. Rate	0.25	0.22	0.23	0.20	0.30	0.14	0.30
(I)	$\widehat{k}$	(2,3,4)	(2,4,5)	(2,3,4)	(2,4,5)	(1,2,3)	(2,2,4)	$= 1$
(II)	Rej. Rate	0.42	0.42	0.43	0.41	0.04	0.93	0.13
(II)	$\widehat{k}$	(3,4,5)	(4,5,5)	(3,4,5)	(3,5,5)	(3,4,5)	(3,3,3)	$= 3$
(III)	Rej. Rate	0.63	0.62	0.64	0.62	0.63	0.37	0.66
(III)	$\widehat{k}$	(2,3,4)	(2,4,5)	(2,3,4)	(2,4,5)	(2,3,5)	(2,3,4)	$= 3$

Table 5.4: $n = 15$ (sample size), $m = 5$ (robustness value). Summary of tests over 5000 repeated simulations of four cases: (0) the null case ($X_i \sim$Exp(1)), (I) a single large outlier ($X_i \sim$Exp(1), $i = 1, ..., 14$; $X_{15} \sim$Norm$(4, 0.1)$), (II) a cluster of multiple outliers ($X_i \sim$Exp(1), $i = 1, ..., 12$; $X_i \sim$Norm$(4, 0.1)$, $i = 13, 14, 15$), (III) multiple dispersed outliers ($X_i \sim$Exp(1), $i = 1, ..., 12$, $X_i \sim$max$(\{X_i : i = 1, ..., 12\})+Exp(1/5)$, $i = 13, 14, 15$). The rejection rate and the median $\widehat{k}$ and quartiles of the estimated number of outliers (in the event of a rejection) are given in alternating rows.

which is exponential ($\alpha = \tau^{-1}$) when $\kappa = 1$, is fat tailed for $\kappa < 1$, and becomes concentrated at τ as κ becomes large. The results of the simulation study are presented in Fig. 5.4 and can be described as follows.

Panel (b) concerns the rate of false positives where $r = 3$ outliers are tested, with level $a = 0.1$, in a Weibull (5.10) sample of size $n = 30$, for a range of shape parameters κ, without outliers. When $\kappa < 1$, the DF is fat tailed, having many events that are large, and thus the tests falsely identify many points as outliers. This is problematic in practice (with small to moderate sample sizes), because one does not know what the true null model is. For instance, with $n = 30$, even when the true DF is considered as an alternative model versus the exponential, and using the powerful LRT, 50 percent of the time (for $\kappa \approx 0.6$), one will not reject the exponential model at a level of 0.1. In this case, when falsely retaining the exponential model, the type 1 error will be between 0.3 and 0.5, depending on the selected TS. The KS test of compatibility of the data with the exponential DF is even less powerful, allowing for more severe false positives.

Case (c) considers the frequency of true positives (power). The setup is the same as above, but 3 dispersed outliers are included. When the Weibull DF becomes less fat tailed, the power of the SRS and MRS tests decreases whereas the power of the D and DK tests increases. Here, with $n = 30$, for the tests of the Weibull versus the exponential, including the outliers in the sample, there is a high probability (0.6-0.8) of not rejecting the exponential model when $1 < \kappa < 1.5$, where the power of some of the tests is weakened.

It is clear that the power, and especially the level, are highly sensitive to the validity of the exponential model, and misspecification of the null can lead to erroneous inference. This has important implications for the practical application of the tests. In particular, one should have a sufficiently large sample to diagnose the validity of the null, and not blindly accept/reject the result of the test and its diagnostics.

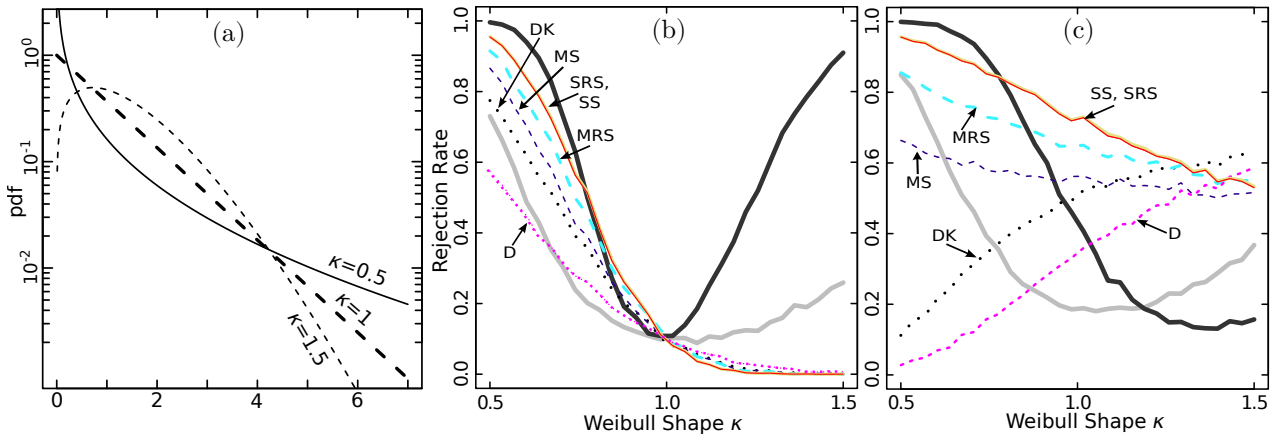

Figure 5.4: **Test robustness** Panel (a): The Weibull PDF (5.10) plotted for parameters (κ, β) equal to $(0.5, 0.4)$, $(1, 1)$ and $(1.5, 1.5)$. Panel (b): The frequency of rejection of the null of no outliers, at level 0.1, in the presense of no outliers, for block tests for $r = 3$ outliers, assuming an exponential null model, when the data is generated from a Weibull for a range of shape parameters κ. Panel (c): The frequency of rejection of the null using a level 0.1, of the block tests for $r = 3$ outliers, with the same setup as frame (b), except that 3 outliers are truly present. The models for the cases are: (b) $X_i \sim$Weibull$(\kappa, 1)$, $i = 1, ..., 30$; (c) $X_i \sim$Weibull$(\kappa, 1)$, $i = 1, ..., 27$, $X_i \sim$max$(\{X_i : i = 1, ..., 27\})$+Exp$(1/3)$, $i = 28, 29, 30)$. For each case, simulation and testing were performed 1000 times for κ sweeping 0.5 to 1.5. The tests are colour coded: SS (red solid), SRS with $m = r$ (yellow solid), MS (blue dashed), MRS with $m = r$ (turquoise heavy dashed), D (magenta light dotted), DK (black dotted). In both frames, the black heavy solid line is the power of the LRT of the Weibull versus the exponential on the data (including outliers). Similarily the grey heavy solid line is for the Kolmogorov-Smirnov test.

5.4 Case studies and "Dragon Kings"

Outlier detection relative to an exponential DF has primarily been motivated by reliability engineering applications. Switching perspective from reliability to risk, the exponential of an exponential variable has the (heavy-tailed) Pareto df (5.2) that is typically used for modeling extremes in both natural and social sciences: earthquake energies, the DF of runs of stock prices, claims in non-life insurance, etc. [79, 183, 189, 243].

The Pareto DF is unique in that it is scale invariant [74, 160], suggesting that events of all sizes – including extremely large ones – are generated by a single mechanism operating at different scales. This feature allows this single parsimonious DF to generate a broad range of event sizes. Thus, if a phenomenon is scale invariant, then extreme events are not predictable and there is nothing anomolous about them as there is nothing to distinguish these events from their smaller siblings, other than their resultant size. This reasoning has been advanced to explain the extreme difficulties in forecasting large earthquakes [102]: according to the approximate scale invariance of the Gutenberg-Richter law, large earthquakes are just earthquakes that started small... and did not stop growing.

However, a number of studies have found either strong or, in other cases, suggestive evidence that there are extreme events "beyond" the Pareto sample [246, 259], i.e., outliers, inspiring the concept of the "Dragon King" (DK) [246] event. DK embody a double metaphor implying that an event is both extremely large (a king [158]), and generated from a unique mechanism/origin (a dragon) relative to other events in the system/sample. The hypothesis advanced in [246, 259] is that DK events are generated by a distinct mechanism (e.g., positive feedback) that intermittently amplifies extreme events, leading to the generation of runaway disasters as well as extraordinary opportunities/successes. Due to the uniqueness of such events, there is hope that such extremes may exhibit precursory signs, disclosing some predictability. The identification of the existence of such phenomena is also clearly important – for example, with applications in risk management. Examples of such DK events have been proposed to include failures of material systems, landslides and some large earthquakes in geophysics, financial crashes in economics, and epileptic seizures and human parturition in biology [246, 259]. Identifying DKs with convincing statistical significance is a prerequisite to the investigation of their origin, understanding their generating mechanisms, and developing forecasting methods, controls, and resilient system designs. Motivated by these considerations, and to provide pedagogical examples, five case studies are considered where DK events are tested as statistical outliers. The case study on

financial market crashes is given below, and the other four in the supplementary material.

5.4.1 Drawdowns/crashes in financial markets

It is well known that crashes in the financial markets occur frequently and can have a significant effect not only on market participants, but also on the broader economy. It is often thought that financial markets are unpredictable – i.e., they are scale invariant / fractal [177, 242] (Pareto distributed). However, in [141, 139, 94] it was found that the sample of crash sizes – measured from the peak to the valley of the event (so-called drawdowns) – contained outliers (defined below). However, the statistical test used in [94] contains an error in the DF of the marginal TS, and [139, 141] did not use standard outlier tests. To correct this, and provide an example, this problem is revisited with the same data. The data are the drawdowns computed for the eleven most actively traded Futures Contracts on the American and European Indices[1], from January 1, 2005 to December 30, 2011.

A peak-to-valley measure of the size of intra-day financial crashes is considered: an *ϵ-drawdown* (hereforth referred to simply as a drawdown) is the total cumulative return of a negative run in price over time, with some specified tolerance for small positive changes along the way [141]. A *drawup* is its positive counterpart. This is an interesting measure of risk because it captures the transient dependence of price changes in time, whereas studying the unconditional df of returns does not. More specifically, considering one trading day $[t_0, t_1]$, prices taken at intervals of width Δ are $p_i = p(t_0 + i\Delta)$, $i = 1, ..., n = \lfloor (t_1 - t_0)/\Delta \rfloor$. The *returns* are then $r_i = \log(p_i/p_{i-1})$. One starts at the first negative return $i_0 = \min\{i : r_i < 0\}$. Then, the cumulative return,

$$r_{i_0,i} = \sum_{j=i_0}^{i} r_j = \log(p_i/p_{i_0}),\ i > i_0\ ,$$

tracks the negative growth of the drawdown, continuing for $i = i_0, i_0 + 1 \ldots$ until the first value of i, say i_2, such that the cumulative return has appeared to reverse direction, relative to its lowest point:

$$r_{i_0,i_2} - \min_{i_0 \le j \le i_2} r_{i_0,j} > \epsilon\sigma\ .$$

Parameter $\epsilon \ge 0$ tunes the tolerance of moves in the opposite direction, and σ is the standard deviation

[1] US: 1) ES, S&P 500, E-mini; 2) NQ, NASDAQ, E-mini ; 3) DJ, Dow Jones, E-mini. European: 4) AEX, Netherlands; 5) CAC, CAC40, France; 6) DAX, Germany; 7) FTSE, UK; 8) IBEX, Spain; 9) OMX, OMX Stockholm 30, Sweden; 10) SMI, Switzerland; 11) STOXX, Euro STOXX, Europe.

of the returns from the previous trading day. The inclusion of σ makes the tolerance adaptive, which allows for volatity regimes. Finally, stepping backwards from i_2, which is the index of a positive change, the drawdown is defined to have occurred from the start i_0 to the lowest point, which occurs at $i_1 = \text{argmin}_{j\in(i_0\leq j\leq i_2)} r_{i_0,j}$. From the next index, $i_1 + 1$, a drawup is defined to begin and computed in a similar way. Drawdowns and drawups alternate in this contiguous way, for the entire trading day.

In panel (a) of Fig. 5.5, for the eight contracts thought to contain an outlier, the largest 5000 drawdowns are plotted according to their empirical CCDF (complementary CDF, i.e., $1 - F(x)$). The empirical CCDF appear approximately linear in the double logarithmic scale, indicating a qualitatively good fit, with the exception of some outliers. There are also some additional differences in the tail. For instance, the tail of the CCDF drops beneath the Pareto fit before crossing back to form the outlying empirical tail. This could suggest an amplification mechanism operating above a threshold size. In panel (c), the Hill plot is given, where the MLE for the tail exponent α is plotted for a range of upper sample sizes. The parameters tend to have an increasing trend, indicating slight convexity in the CCDF in panel (a), and thus a loss of outlier testing power for large sample sizes (Sec. 5.3.4). Based on the Hill plot, the estimator for the top 1'000 points appears to be approximately stable for most of the contracts. For systematic threshold / sample fraction selection, three methods are used: 1. comparing the AIC of exponential and nonparametric (R:logspline) fits, 2. selecting the smallest KSD as recommended in [52], and 3. selecting the smallest threshold where the KS test p-value is above 0.10. The results of these methods are given in Tab. 5.5. All but one of the 24 tests select at least the top 1'000 points, thus upper samples of this size and smaller will be considered for outlier testing.

Test	ES	CAC	DAX	FTSE	SMI	IBEX	NQ	DJ
AIC	1'184	1'214	2'290	2'734	2'704	2'055	1'154	3'757
KSD	1'049	1'115	1'520	3'144	609	1'501	1'074	1'134
p	1'985	1'323	2'403	3'714	2'701	2'255	3'123	4'000

Table 5.5: The selected number of points in the upper sample based on comparing the AIC of the null and the nonparametric model, minimizing the KSD, and the largest upper sample with KS test p value greater than 0.1.

In panel (a) of Fig. 5.5, the apparent outliers are large and dispersed. Thus, the MRS TS (5.3) should be powerful (Sec. 5.3.3) and can be applied inward for a range of thresholds, requiring a fraction of the computation of outward testing. For each dataset, the inward test was performed – with MRS TS, $m = 10$, level $a = 0.1$, and upper sample size ranging from $n = 10$ to $n = 1000$. For all contracts,

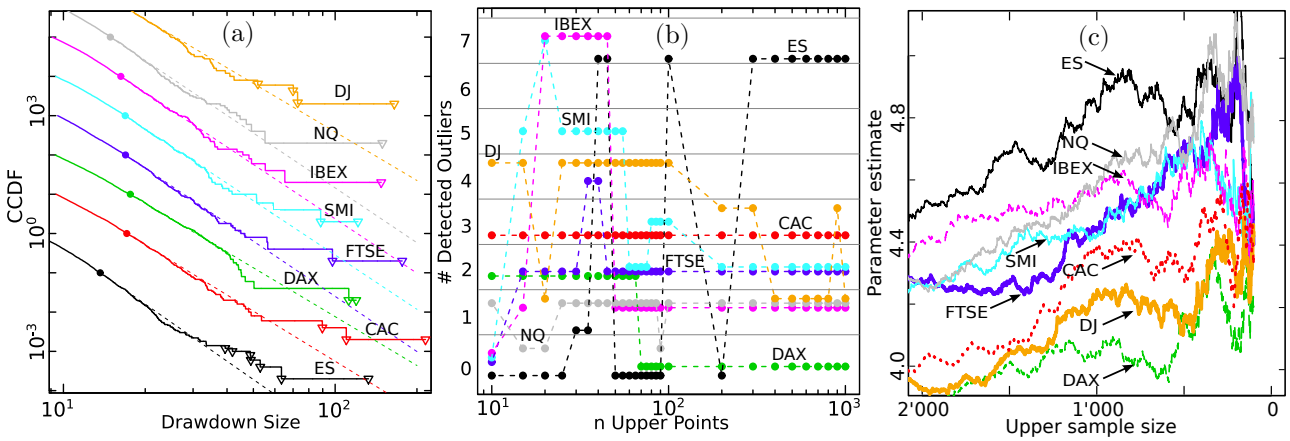

Figure 5.5: **Financial Market Crashes. (a):** The 5000 largest drawdowns for each of the 8 futures contracts thought to contain outliers, plotted according to their empirical CCDF in double logarithmic scale. For clarity, each CCDF above the black one is multiplied by 10 relative to the one beneath it. The Pareto DF with MLE parameter for the top 500 points is given by the dashed lines, starting at the solid dot. The triangles identify the points that were identified as outliers based on the interpretation of panel (b). **(b)** The number of identified outliers is plotted against sample size where the MRS test (5.3) with level $a = 0.1$ has been applied inward with $m = r = 10$, for a range of sample sizes n, for each contract in (a) with the same colour coding. **(c)** Hill plot: The estimated tail exponent is plotted for a range of upper sample sizes. (see online version for colour)

excluding AEX, OMX, and STOXX, at least 1 outlier was found and are indicated in Panel (b) of Fig. 5.5. For some of the contracts, the results are quite stable across sample size (e.g., CAC and FTSE). For others, the impurity of the DF plays a role in the interpretation. For instance, for DAX, two outliers are detected once the test is restricted to the bent-down tail. For ES, choosing between zero and seven outliers is more subjective – are there multiple outliers, or does the tail grow heavier? For IBEX, it is clear that the identification of seven outliers is due to the dip in the empirical CCDF occuring between drawdown size of twenty and thirty. The alternative choice of 1 outlier is more stable with respect to a broad range of values of n. The interpreted outliers are indicated in panel (a).

The largest outliers coincide with major news events: The 07 July 2005 London bombings coincided with the largest outliers of CAC, DAX, FTSE, SMI, and IBEX – all being based on European indices. Further, DAX and CAC each have an outlier corresponding to the "Mini Flash Crash" of 27 Dec. 2010 (e.g., see [41]). All American contracts (ES, DJ, and NQ) have their largest outliers coinciding with the infamous "2010 Flash Crash" of 6 May 2010. We thus observe that outliers occur either due to some exogenous impacts (London bombings) or as a result of an endogenous transiently

unstable dynamics (flash crash). Indeed, in [91, 90], it was suggested that financial markets exhibit a significant endogeneity or "reflexivity", in the sense that nowadays up to 70-80% of trades occurring at the time scales of fractions of seconds to tens of minutes are motivated (or triggered) by previous trades. In this framework [91, 90], dragon kings emerge when the market dynamics become critical and super-critical, that is when the future trades are triggered only by previous trades and not by news, making the financial markets essentially self-referential in these periods. Thus, some of the outliers can be classified as dragon king drawdowns.

5.4.2 Nuclear accidents

We consider as events accidents occurring at nuclear power plants, studied in [291]. For this two measures of severity are considered: the cost measured in 2011 US Dollars, for which there are 173 values over the period of 1960 to 2015; and a logarithmic measure of radiation released called the Nuclear Accident Magnitude Scale (NAMS) [236], for which there are 33 values over the same period. Since the disaster at Fukushima in 2011, Nuclear power has come under major public scrutiny. Further, the level of risk that the nuclear industry claims is consistently much lower than statistical analysis of past events indicates [256]. Thus, it is crucial to arrive at a better understanding of the true risk level in this critical application.

The disasters occurring at Chernobyl (1986) and Fukushima (2011) are the most costly accidents thus far, and together are estimated to have caused damage costing 430 Billion 2011 US dollars. This is roughly equal to five times the cost of all 173 other events together. These events, together with TMI (Three Mile Island, 1979), are also the largest radiation release events. These events are thus extremely large. It is instructive to ask whether a heavy Pareto tail is sufficient to account for these extreme risks or, alternatively, if the tests discussed here can identify outliers / DKs in this data.

In Fig. 5.6 the empirical CCDF (complementary CDF i.e., $Pr\{X > x\}$) for NAMS and the log cost are plotted. For log cost only the 114 events occurring post TMI and included due to an abrubt change in distribution after TMI. For NAMS, the largest three events form a cluster, and appear outlying relative to the Exponential df with $\hat{\alpha}_{NAMS} = 0.7$ (0.3) fit by MLE to the top 15 points. Not surprisingly the df of NAMS and the log cost are similar, as they are certainly related. For log cost, the two or three largest events appear to be outlying relative to the Exponential df with $\hat{\alpha}_{\$} = 0.6$ (0.14) fit by MLE to the top 50 points. As shown in the hill plot, inset in Fig. 5.6, when comparing the

AIC of the logspline nonparametric fit with the exponential one, the exponential cannot be rejected for samples smaller than the 60 largest points. Further, when performing the KS test, the exponential fit cannot be rejected (at a level of 0.05) for samples smaller than the 80 largest points. Thus the exponential approximation for the tail, and thus the outlier test, should only be applied to not more than the upper 60 points.

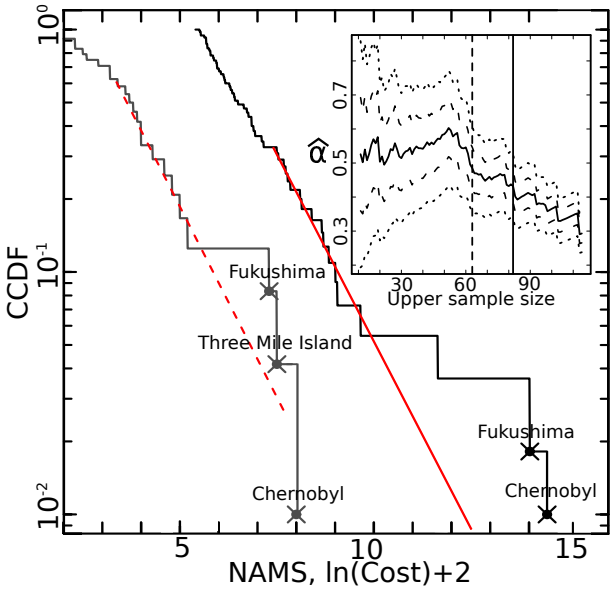

Figure 5.6: **Nuclear Power Plant Accidents:** the CCDF (cumulative complementary distribution function, i.e., $Pr\{X > x\}$) of the log of the 56 largest cost events, in millions of US Dollars, (black solid) shifted by 2 units for visibility, and the CCDF of all 33 log radiation release (NAMS) values, in solid grey. The fitted lines are exponential MLE fits. The inner panel is the Hill plot for the cost values. The solid rough line is the MLE for the exponential distribution for the tail of cost events, for multiple upper sample sizes. It is bracketed by lines indicating one and two standard deviations of the estimator. The vertical dashed line indicates the largest sample at which the exponential cannot be rejected (based on AIC) as being as good as the logspline nonparametric fit. The vertical solid line indicates the largest sample at which the exponential cannot be rejected by the KS test.

We now test the outliers with a number of the aforementioned tests. The results are presented in Tab. 5.6 and summarized below. First considering NAMS, in Fig. 5.6 the CCDF is visibly concave until the top 15 points or so, causing a decrease of test power for tests applied to larger sample fractions (Sec. 2.7). Since the outliers are clustered, (from Sections 2.4 and 2.6) the mixture approach is most powerful, and inward tests the weakest. Despite the small sample size, the mixture test consistently identifies 2 or 3 outliers over a range of upper samples. This confirms that the cluster of large events is a significant feature, however this cluster of large events is not far enough beyond the tail for the

other tests to reject the null. It is also important to note that the sample size is very small, and thus our ability to diagnose the validity of the null is weak! Next, cost values are considered for which a larger sample is available. The outward test and the mixture test consistently identify the two largest points as significant outliers. The SRS block test fluctuates around a value of about 0.1. It is not surprising that tests based on the MRS fail to reject due to lack of power when the largest point is not extremely large. Thus Fukushima and Chernobyl appear to be outliers in both radiation released (NAMS) and cost. This is compatible with our understanding of these accidents, where the disaster escalated beyond the threshold of control, leading to an unmitigated proliferation of damage. That these points are outying in both (dependent) samples would give higher significance if a bi-variate outlier test were performed.

It is worth mentioning that there is a positive relationship between NAMS and cost: Considering the 30 events with substantial radiation release (NAMS>0), a linear regression of the logarithm of cost (the response) versus NAMS (the explanatory variable) yields an intercept of 2.33 (0.7), $p = 0.003$ and a slope of 0.97 (0.17), $p < 10^{-5}$, with coefficient of determination $R^2 = 0.5$. Further, the same regression can be done for the 16 events that have occurred at Sellafield, in the UK. The result of this is an intercept of 2.30 (1.0), $p = 0.04$ and a slope of 1.17 (0.39), $p = 0.001$, with coefficient of determination $R^2 = 0.4$. Thus, there is a significant relationship between radiation release and cost, where we have simply considered a linear relationship. Of course the regression parameters for different plants will depend on the value of property development around the plant.

Data	n	$r = m$	MRS	SRS	MS Out	MRS In	Mix	DK
NAMS	20	3	0.62	0.35	0, $0.09 > 0.04$	0, 0.62	$\mathbf{2, 0.03}$	0.21
NAMS	15	3	0.60	0.32	0, $0.07 > 0.04$	0, 0.60	$\mathbf{2, 0.025}$	0.20
NAMS	10	3	0.37	0.15	$\mathbf{3},\ \mathbf{0.025 < 0.04}$	0, 0.37	$\mathbf{3, 0.025}$	0.14
Damage	50	2	0.17	**0.08**	$\mathbf{2, 0.03 < 0.06}$	0, 0.17	$\mathbf{2, 0.05}$	0.18
Damage	40	2	0.23	0.11	$\mathbf{2, 0.04 < 0.06}$	0, 0.23	$\mathbf{2, 0.06}$	0.22
Damage	20	2	0.25	0.14	$\mathbf{2},\ \mathbf{0.05 < 0.055}$	0, 0.25	$\mathbf{2, 0.07}$	0.25
Damage	15	2	0.17	**0.07**	$\mathbf{2},\ \mathbf{0.02 < 0.04}$	0, 0.17	$\mathbf{2, 0.03}$	0.21
Damage	10	2	**0.06**	**0.02**	$\mathbf{2},\ \mathbf{0.01 < 0.04}$	$\mathbf{2},\ \mathbf{0.01}$	0, 0.18	0.16

Table 5.6: Summary of outlier tests for NAMS and cost data for the upper n points, for r outliers (with robustness value $m = r$). Bold values indicate significance at a level of $a = 0.1$. Block tests performed include: MRS (7), SRS (5), mixture likelihood ratio (10), and the DK test (9). Further, the MS (6) test was applied outward (MS Out), with the number of identified outliers, the p-value, and the adjusted level (to achieve $a = 0.1$) given. For instance, in the first row for MS Out there are zero outliers because the p-value of 0.09 is above the adjusted level of 0.04. Finally, the MRS test was applied inward (MRS In), with the number of identified outliers, and the p-value of the test for the largest point given.

5.4.3 Stock returns

An issue of debate is if the 1987 stock market crash (Black Monday) was an outlier. We focus on [224], which is the most recent study on this problem. In [224], considering daily returns on the Dow Jones Industrial Index, from 3 January 1977 to 31 January 2005, it was claimed that Black Monday is not an outlier. In further detail, the returns were whitened by taking the residuals of a standard AR(1)-GARCH(1,1) model estimated on the returns. Next, the two largest whitened returns $X_{(2)}$ and $X_{(1)}$ were tested as outlying. The test used relies on the GPD approximation (2) of the tail of the sample, and requires an estimate of the tail parameter α. A sample size of $n = 732$ was used to estimate α. The test statistic $T_r = X_{(r)}/X_{(r+1)}$, comparing $X_{(r)}$ to the previous (next largest) order statistic $X_{(r+1)}$, was used to test if $X_{(2)}$ and $X_{(1)}$ were outlying. Testing outward, with a level of 0.05, neither of these points were identified as outliers.

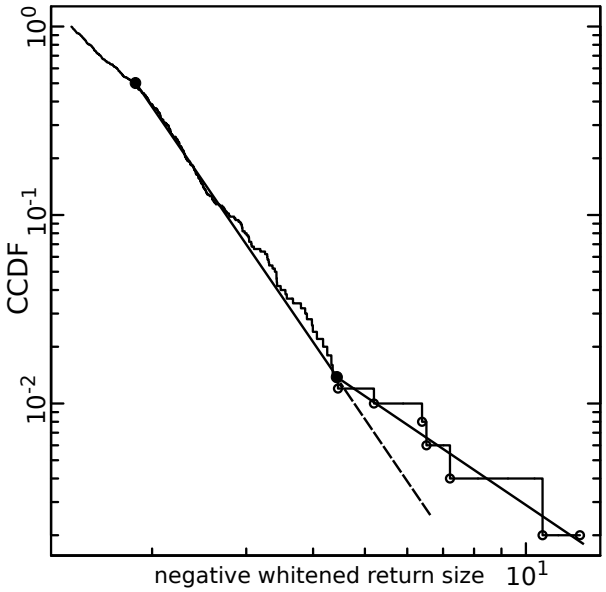

Figure 5.7: **Stock Returns:** The rough line provides the empirical CCDF of the magnitude of the 500 largest whitened returns of the Dow Jones Industrial Index from 3 January 1977 to 31 January 2005. The solid lines between solid dots provide Pareto model estimates for two magnitude layers. The dashed line extends the slope of the first layer for comparison with that of the second.

To evaluate the approach taken in [224], we first plot the CCDF of the 500 largest negative whitened returns in Fig. 5.7. This plot was not provided in [224], but is clearly essential to assessing above which threshold the GPD approximation (2) is sound. A few important points are apparent from the figure: Firstly, the CCDF above the 200 largest observations is shallow/concave, and thus

considering more than 200 points (i.e., 732 in [224]) in the sample will weaken the test (i.e., the estimate of α will be too small). Secondly, the second largest point is similar in magnitude to the largest. Thus, the test using $T_1 = x_{(1)}/x_{(2)}$ will be masked by $x_{(2)}$, and not rejected. Finally, the top 6 or 7 points seem to follow a heavier tailed df. Thus, 6 or 7 points should be tested as outlying, rather than only 2, and a sum test statistic, measuring the cumulative departure of the empirical tail, could be more powerful.

First, we consider estimating a Pareto df with two layers. The first layer, containing 193 points, covers $1.97 < x \leq 4.45$ and has MLE $\hat{\alpha}_1 = 3.8$. The second layer, containing the 7 largest points, covers $4.45 < X$ and has MLE $\hat{\alpha}_2 = 1.8$. Given that the first layer model is true, there is a $p = 0.02$ probability of observing such an extreme difference between the estimated parameters. This two layer model appears to describe the empirical CCDF well (Fig. 5.7). Next, a single layer model for the top 200 points, covering $4.45 < X$ was estimated with MLE $\hat{\alpha}_0 = 3.9$. The likelihood ratio test of the two layer versus one layer model is rejected in favour of the two layer with p-value 0.07. Further, applying the SS test for $r = 6$ with the top 200 points rejects that there are no outliers with $p = 0.04$. Finally, applying the DK test for 6 outliers, for upper sample sizes ranging from 20 to 200, all tests had $p < 0.04$. Thus it appears that the 6 largest points are outlying.

The largest one is, unsurprisingly, "black monday" Oct. 19, 1987, which is unambiguously classified as an outlier. An enormous literature has dwelled on its possible origin with a lot of confusion as no simple proximate cause can explain its occurrence. We find more compelling the story that it marked the end of a large financial bubble and thus corresponded to its burst [254, 242, 142]. The second largest event occurred on "black friday" Oct. 13, 1989 and is usually associated with a fall of the junk bond market (`https://en.wikipedia.org/wiki/Friday_the_13th_mini-crash`). The third largest loss corresponds to the first day of reopening of the US stock markets on Sept. 17, 2001 after Sept. 11, 2001. It is not clear to us how to interpret the fourth largest loss that happened on Nov. 15, 1991. The fifth largest loss on Oct. 27, 1997 is analyzed in details in [242], which paints a picture much richer than the usual story that this was a global stock market crash caused by an economic crisis in Asia. This loss can actually be seen also as a partial burst of a bubble that had been surging in the few previous years (recall the famous quip on the "irrational exuberance" of the stock markets by Alan Greenspan, then the Chairman of the US Federal Reserve, on Dec. 5, 1996 (`http://www.federalreserve.gov/boarddocs/speeches/1996/19961205.htm`)). The sixth largest

loss on Nov. 9, 1986 is not clearly associated with any exogenous cause, to the best of our knowledge. These six outliers are part of the list found by other researchers (e.g. [98]).

5.4.4 Fatalities in Epidemics

We now study the number of fatalities caused by outbreaks of bacterial, viral, and parasitic diseases (epidemics). A dataset for this, with 1,368 events covering the period from 1900 to 2015, was provided by [106]. The dataset excludes, and in some case provides only national fatalities for, pandemic events (spanning multiple countries). Thus the dataset was complemented with the well known Spanish (1918), Asian (1957), and Hong Kong (1968) Influenza pandemics, which each caused in excess of 1 million fatalities [203]. Further, the 2009 H1N1 "Swine" influenza pandemic, which was estimated to cause upwards of 150,000 fatalities [234], was also included. All epidemics and pandemics will be simply referred to as events.

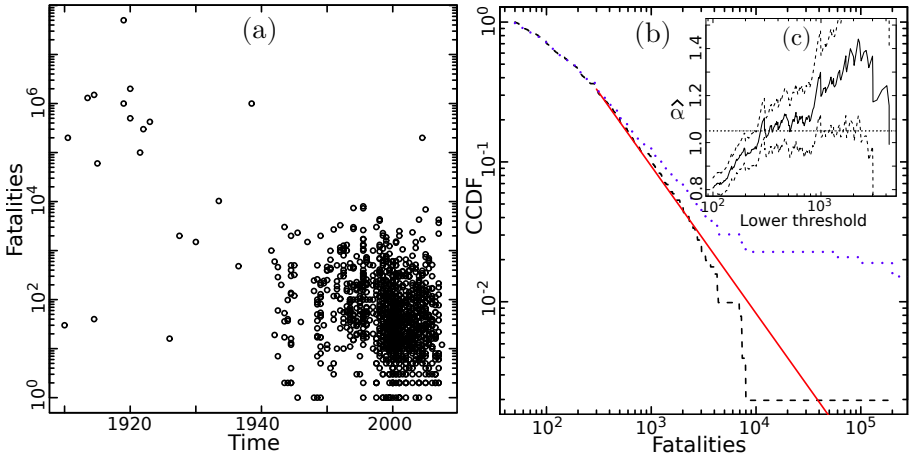

Figure 5.8: *Epidemic Fatalities:* **(a)** scatterplot of 1,368 epidemic fatalities from 1900 to 2015. **(b)** The CCDF of the 507 events in excess of 50 fatalities from 1960-2015 (black dashed), its Pareto tail fit with lower threshold $u = 300$, and MLE $\hat{\alpha} = 1.05$ (0.08), and the CCDF of the 523 events in excess of 50 fatalities from 1900-2015 (blue dotted), having 9 events beyond the window. The inner panel **(c)** plots the Pareto tail estimate for samples exceeding lower thresholds ranging from 50 ($n = 507$) to 4,300 ($n = 20$) for the time period from 1960-2015 (i.e., the black dashed CCDF). The level 1.05 is given by the horizontal line.

From Panel (a) of Fig. 5.8, it is clear that over time the dataset has become more complete, in particular for small event sizes. Further, in the period from 1900-1960, 13 events have more than 10,000 fatalities (0.21 per year), whereas in the period from 1960-2015, only 1 such event does (0.02 per year). Notwithstanding potential changes in the true frequency of events, this is obviously a highly significant difference. These historical extreme events – Influenzas, Bubonic plagues, Cholera, etc. – have largely been eradicated through sanitation, vaccines and antibiotics.

Considering the period from 1900 onwards, many changes have occurred that should have influenced both the incidence and severity of events. Due to data incompleteness, the rate of events cannot be studied.

Despite this, the sample in excess of 50 fatalities from 1960 onwards, containing 507 points, is roughly stationary in severity. For instance, when repeatedly (1000 times) sampling 100 points from the 507 points, splitting the 100 points into two equal subsamples, and testing their distributions for equivalence with the KS test, only 12.6 percent of p-values were less than 0.1. Thus, the modern sample – spanning the 55 years following 1960 – may be used as a proxy to evaluate the outlyingness of the historical extremes, or at least to evaluate how outlying they would be if they were to occur now.

The events in excess of 50 fatalities from both 1900 onwards and 1960 onwards are plotted according to their CCDF in Panel (b) of Fig. 5.8. The sample from 1960 approximately has a Pareto tail (see Panel (c)) with parameter around 1.05 (0.08) for the 168 points above the lower threshold of 300. With increasing lower truncations, the estimated parameter increases (as the CCDF bends down), however this is not a significant departure from the estimated tail. For instance, the Anderson-Darling test for the fit of the top 168 points gives a p-value of 0.8. The tail of the sample from 1900 onwards is skewed both by the inclusion of historic large events, and also by the absence of their smaller siblings, which were not recorded.

The value of the exponent $\alpha \approx 1$ is reminiscent of Zipf's law, which is known to derive quite robustly from the interplay between three simple ingredients [216]: birth, proportional growth (also known as "preferential attachment" in network theory) and death. If the variance of the proportional growth component is large, the df of event sizes converges to a power law with exponent $\alpha \approx 1$. These ingredients are arguably minimum constituents of epidemic processes and rationalize our finding $\alpha = 1.05$ (0.08). What is really surprising is the detection of outliers that we present below, which, in some cases, suggests the activation strong amplification processes beyond the proportional growth mechanisms.

We turn our attention to the detection of outliers relative to the approximately stationary data from 1960 onwards. The 14 events in excess of 10,000 fatalities – 13 of which happened before 1960 – are considered. The smallest of these 14 is a Cholera outbreak causing 10,276 fatalities (Egypt, 1947). We start with the weakest possible test, considering as a sample: the 167 points with between

300 and 10,000 fatalities occuring since 1960, plus the aforementioned Cholera outbreak. Testing for a single outlier with the DK test (9) gives a p-value of 0.002. Thus any of the other suspected outliers – including the 2011 Swine Flu event – would be identified as significant outliers also. And, including multiple of these outliers in the sample, and testing them together, would provide even higher significance.

With respect to the mechanism(s) at the origin of these outliers, it is likely that each case may be associated with specific catalysing processes. For one of the largest dragon-kings, the so-called Spanish flu of 1918 which killed an estimated 50 millions people in the world, there is a clear identified amplification mechanism. In this epidemic, about 500–600 million people, a third of the world's population at that time, were infected. The pandemic took five times more lives than the First World War. The first cases of the unknown disease were registered in Kansas, America, in January 1918. By March 1918, more than 100 soldiers fell ill at the US army camp in Funston, Haskell County, where more than 5000 recruits were training for further military operations on the European battlefronts of the First World War. Most of the recruits were farmers, had regular contact with domestic animals and were less resistant to viruses than recruits from cities. The high concentration of personnel in the camp simplified human-to-human transmission. At that time, viruses were not known to medicine, and some doctors had not even accepted the idea that microorganisms could cause disease. Later, the personnel of Funston camp were transferred to Europe by ship, and during the long transatlantic crossing, the virus spread among soldiers coming from other parts of the USA. Upon arriving in Europe, American soldiers infected British and French forces, which in their turn infected German forces in hand-to-hand combat. When Woodrow Wilson, President of the United States from 1913 to 1921, began to receive reports about a severe epidemic among American forces, he made no public acknowledgement of the disease [27]. Moreover, other governments involved in the war made similar decisions – censorship, lies, and even active propaganda – to keep up morale, allowing the disease to continue to spread without any preventive measures. The pandemic was named "Spanish flu" because Spain was a neutral country during the First World War and did not suppress the media, so it was only Spanish newspapers that published honest articles about the severity of the disease – despite the fact that it had originated in the USA and spread initially among American soldiers in the absence of a proper response by the US government. This lack of response was probably due to the US strategic goal of developing a strong political influence in the post-WWI peace process that was to shape international politics in the

following decades. In summary, the amplification mechanisms that led to the Spanish flu dragon-king are (i) extremely efficient connectivity between people mediated by movements of soldiers and (ii) rare absence of any prophylactic or treatment measures due to the priority given to the war efforts.

We thus conclude that we found evidence of dragon-kings in the database of epidemic events, including the more recent period post-1960, albeit with a much reduced frequency. For instance, one of our detected outliers, the Swine Influenza pandemic, occurred in 2009. Concerning the AIDS pandemic, which is not included in the dataset, in 2014, 1.2 million [1 million–1.5 million] people died from AIDS-related illnesses, a significant improvement from the maximum reached in 2015 of 2.3 million [2.1 million–2.6 million] deaths from AIDS-related illnesses, with an estimated $\sim$ 36 million total deaths since its identification [277, 276]. The evidence we have presented for a dragon-king regime in the dynamics of epidemics suggests that a return of pandemic plagues cannot be ruled out, perhaps catalysed by the severe progressive threats of antimicrobial resistance [53], and climate change.

5.4.5 City sizes

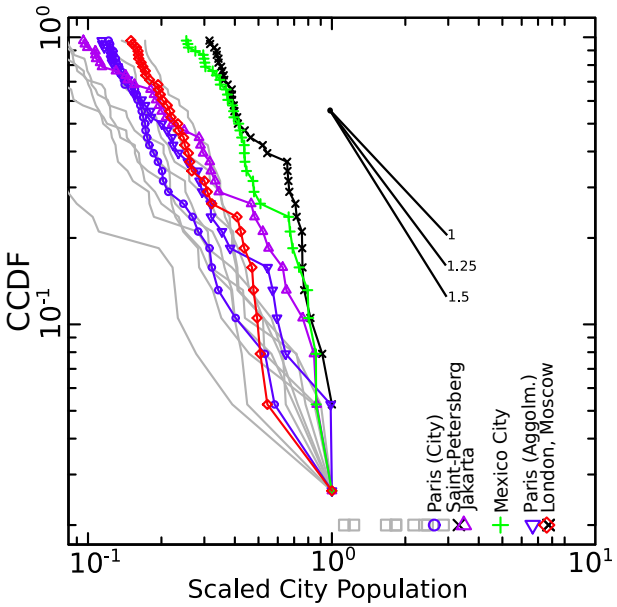

Figure 5.9: **City sizes**: plot of the CCDF for the 35 largest cities (and also agglomerations for France) in each of the 14 countries: Brasil, China, France, India, Indonesia, Japan, Korea, Mexico, Nigeria, Pakistan, Phillipines, Russia, the UK, and the USA. The sizes were scaled such that the second largest point (third largest for Russia) in each country has size 1. The scaled largest point (two largest for Russia) are plotted in the bottom right. Each country that is suspected of having outliers is in colour: France (blue circles for cities, blue downward triangles for agglomerations), Russia (black x marks), Indonesia (purple triangles), Mexico (green crosses), and England (red squares).

Within the disciplines of economics, geography and geopolitics (among others), the distribution of city and of agglomeration sizes is of particular interest, due to the importance of urban primacy, and because it constitutes one of the key stylized facts. There is a large literature documenting that the distributions of city and agglomeration sizes follows a Pareto df with parameter close to one (Zipf's Law) (see e.g. [216] and references therein). There has been some debate over if the df would be better represented by a Lognormal [75, 76, 162], however the debate has been clearly settled in favour of the Pareto for the 1000 largest cities [175]. Note that both the Pareto and Lognormal df's are generally taken to result from Gibrat's principle of proportional growth [104] (see [216] for a general derivation).

In [201], the DK test (9) was used to identify outlying population agglomerations for a number of countries, assuming a Pareto tail. Here we consider city sizes rather than agglomerations since this data is available for more countries. We only consider agglomeration sizes for the case of Paris, France

for comparison with [201]. Data for 14 large countries[2] were taken from [1]. All tests use the SRS block test statistic for testing the largest point as an outlier, with the exception or Russia where two outliers are tested.

In Fig. 5.9, the 35 largest cities of each country are plotted according to their empirical CCDF, rescaled in a way to make the largest cities comparable. Since not all of the samples appear to follow a pure Pareto df, results on robustness and testing the tail (Sections. 2.1 and 2.7) are relevant here. First considering French cities, for upper sample sizes of $5 < n \leq 35$, the p-value fluctuates in a range of $0.1 - 0.2$. Thus, there is only marginal evidence that the city of Paris is an outlier. However, the agglomeration of Paris is relatively larger, and for $5 < n \leq 25$ the p-value fluctuates between 0.02 and 0.15, providing stronger evidence of the uniqueness of Paris. The CCDF of Indonesia is concave. Thus, if too large of a sample is considered in the test, Jakarta will not be detected as an outlier. For instance, if one draws a line that best interpolates all points of the empirical CCDF, the line will be so shallow that the Jakarta point falls beneath it, essentially masking the outlier. For this reason, Jakarta, Indonesia has $p < 0.1$ only for the upper most points $5 < n < 11$. Mexico is an even more extreme case of the above, having $p < 0.1$ for $5 < n < 20$ for Mexico City. London, UK, is the most significant, having $0.001 < p < 0.05$ for all $5 < n \leq 35$. Finally, testing both Moscow and Saint-Petersberg as outliers, the p-value is in $0.01 < p < 0.15$, with a mean of 0.09 for all $5 < n \leq 35$. In conclusion, it is absolutely clear that London is an outlier, and the largest city/cities of five of the remaining fourteen countries considered have moderate/suggestive evidence that they are outlying.

5.5 Discussion

We provided a comprehensive study of outlier detection in the highly general case of samples with exponential and Pareto tails. By considering a variety of TS and outlier scenarios, many useful insights are provided for practitioners. Further, a simple yet novel modification of TS was shown to make the convenient inward test competitive with the relatively arduous outward test.

Insights include that one should select the correct TS based on the the nature of the suspected outliers. For instance, a mixture model can be very useful for clustered outliers, whereas an inward test with a MS type statistic will be powerless. Next, the power and level of outlier tests are highly

[2]Brasil, China, France, India, Indonesia, Japan, Korea, Mexico, Nigeria, Pakistan, Phillipines, Russia, the UK, and the USA.

sensitive to the correct specification of the main DF (exponential/Pareto). For robust results, it may be better to focus on the tail of the sample, where EVT provides that the best approximation is attained. If the approximation is poor even in the tail, then one should choose a better null model to avoid spurious inference. Further, tests should be applied for a range of upper samples sample (growing lower threshold) and consistent rejection required for a robust rejection to be verified.

In the case studies, the concept of Dragon King events was introduced. This stresses that some outliers are meaningful, and perhaps special. Further, one should certainly not simply discard these outliers but rather focus on understanding them. Significant outliers were found in the sizes of financial returns and crashes, epidemic fatalities, nuclear power generation accidents, and city sizes within countries. In the cases of financial crashes and nuclear accidents, the existence of dragon kings should be considered in the assessment of risk.

Part III

Studies of extreme risks

Chapter 6

Nuclear risk

This chapter is based on [291].

6.1 Introduction

The industry-standard approach to the evaluation of the risk of nuclear accidents is a top-down technique called probabilistic safety assessment (PSA). PSA consists of developing fault tree models that allow one to simulate accidents, with different triggers and event paths, and the severity and frequency of such accidents [159]. Furthermore, within a plant, PSA may be an ongoing process where both the PSA, and plant operations and technology, evolve together with the purpose of improving plant safety. The basic PSA methodology works as follows: Initiating events, such as component failures, human errors, and external events, are enumerated and assigned probabilities. Next, a (typically deterministic) fault tree is defined to encode the causal links between events, allowing combinations of initiating events to form the ultimate resultant/system-level event. Such a model then allows one to determine the probability of such events, and potentially attach damage/consequence values to the event paths. Thus, a textbook PSA would require the complete and correct definition of initiating events, subsequent cascade effects, and their probabilities and consequences.

It is therefore not surprising that the documentation for a plant-specific PSA often fills a bookshelf, and is a constant "work in progress". Within PSA, three levels exist, delineating the depth/extent to which events are studied [129, 130, 133]: level 1 concerns core damage events, level 2 concerns radioactive releases from the reactor building given that an accident has occurred, and level 3 evaluates the impact of such releases on the public and the environment. Levels 1 and 2 are required by reg-

ulation. Level 3, which is the level considered in this study, is seldom done in PSA. Given that the reliability of PSA depends on the inclusiveness of scenarios, the correct modeling of cascade effects, and the handling of tremendous uncertainties, it is not surprising that PSA has failed to anticipate a number of historical accidents in civil nuclear energy [257, 156]. In [172], it was found that the probability assessments were fraught with unrealistic assumptions, severely underestimating the probability of accidents. In [5], the chairman of the World Association of Nuclear Operators stated that the nuclear industry is overconfident when evaluating risk and that the severity of accidents are often underreported.

Instead of entering this quagmire, several studies have used a "bottom-up" approach, performing statistical analysis of historical data. These studies [257, 126, 236, 108, 85] and others have almost universally found that PSA dramatically underestimates the risk of accidents. The IAEA (International Atomic Energy Agency) provides the INES (International Nuclear Event Scale) measure of accident severity, which is the standard scale used to measure the severity of nuclear accidents. However, the INES has been censured – for being crude, inconsistent, only available for a small number of events, etc. – not only in statistical studies, but by the industry itself [5, 40]. As noted by The Guardian newspaper, it is indeed remarkable (sic. astonishing) that the IAEA does not publish a historical database of INES events [214]. However, given that the IAEA has the dual objective of promoting and regulating the use of nuclear energy, one should not take the full objectivity of the INES data for granted. Independent studies are necessary to avoid possible conflicts of interest associated with misaligned incentives.

Presumably for lack of better data sources, a number of statistical studies such as [108, 85] have used the INES data to make statements about both the severity and frequency of accidents in nuclear energy systems. Here, we also perform a statistical analysis of nuclear incidents and accidents, but we avoid relying on the INES data. Instead, we use the estimated cost value in USD (US dollars) as the common metric that allows one to compare often very different types of incidents. This database has over triple the number of events compared with most studies, providing a much better basis for statistical analysis and inference, and bringing into question the reliability of the other studies. Moreover, because radiation releases may translate into very different levels and spread of contamination of the biosphere depending on local circumstances, the quantification of cost is more useful and provides a better comparative tool.

According to PSA specialists, the gaps between PSA-specific results and the global statistical data analysis mentioned above exist in the eyes of observers who ignore the limitations in scope that apply to almost all PSA – e.g., PSA is often restricted to normal operating conditions and internal initiating events. Indeed PSA is a tool that serves many purposes that do not rely on the accurate absolute quantification of risks. However, PSA *is* used as the tool for discussing risks in nuclear energy systems, and has multiple shortcomings in this regard. That is, PSA applications need to better consider incompleteness, uncertainty [186, 156, 208, 88], and be combined with bottom-up statistical approaches when discussing risks at many levels [171].

Moreover, because of the uniqueness of each reactor, some nuclear experts say that assigning risk to a particular nuclear power plant is impossible [131, 2]. A further argument is that the series of accidents form a non-stationary series in particular because the industry has been continuously learning from past accidents, implementing procedures and upgrading each time to fix the problem when a vulnerability was found especially via accidents. For instance [157]: the loss of criticality control in the fast breeder reactor EBR-I (1.7MWe) that started operation in 1951 on a test site in the Idaho desert led to a mandatory reactor design principle to always provide a negative power coefficient of reactivity when a reactor is producing power; the Windscale accident in 1957 catalyzed the establishment of the general concept of multiple barriers to prevent radioactive releases; the Three Mile Island accident in 1979 led to plant specific full-scope control room simulators, plant specific PSA models for finding and eliminating risks and new sets of emergency operating instructions; the Chernobyl accident in 1987 led to the creation of the World Association of Nuclear Operators (WANO) through which participating operators exchange lessons learned, and best practices; the Fukushima-Daiichi accident in 2011 is pushing towards designs that ensure heat removal without any AC power for extended times., etc. As a consequence, each new accident supposedly occurs at a nuclear plant that is not exactly the same as for the previous accident. This leads to the concept that nuclear risks are unknowable because one does not have a fixed reference frame to establish reliable statistics [4].

In contrast, we propose that it is appropriate – and important – to study the global risk of nuclear energy systems by performing statistical analysis of an independently compiled dataset of the global pool of reactors. There is nothing invalid about modeling the overall risk of the heterogeneous global fleet, provided one takes sufficient care to control for non-stationarity, and does not draw inference beyond the "average reactor". In particular, risk-theoretic stochastic models aiming at describing both

the frequency and severity of events, as in [126, 257], offer very useful guidelines for such statistical analyses. This constitutes the standard approach that insurance companies rely upon when quoting prices to cover the risk of their clients, even when the estimation of risk appears very difficult and non-stationary. In this spirit, Burgherr et al. [42] write that "the comparative assessment of accident risks is a key component in a holistic evaluation of energy security aspects and sustainability performance associated with our current and future energy system."

In the next section, we describe the data used in our analyses, how severity of events in nuclear energy systems can be measured, and show that the INES values are a poor measure of severity when compared with the consequences of events measured in USD cost. Section 3 discusses uncertainty quantification in nuclear risks. Section 4 estimates the rate of events and proposes simple models to account for the evolution of the nuclear plant industry. Section 5 analyses the distribution of costs. Section 6 discusses a runaway disaster effect where the largest events are outliers referred to as "dragon-kings" (DK). Section 7 combines the different empirical analyses of previous sections on the rate of events, the severity distribution and the identification of the DK regime to model the total future cost distribution and to determine the expected annual cost. Section 8 concludes and discusses policy implications.

6.2 Data and the measurement of event severity

We define an "event" as an incident or accident within the nuclear energy system that had material relevance to safety, caused property damage, or resulted in human harm. The nuclear energy system includes nuclear power plants, as well as the facilities used in its fuel cycle (uranium mines, transportation by truck or pipeline, enrichment facilities, manufacturing plants, disposal facilities, etc.). Events are defined to be independent in the sense that one event does not trigger another one. For instance, three reactors melted down at Fukushima in 2011, however we define this as a single accident due to the fact that the occurrences at the individual reactors were part of a dependent sequence, and linked to a common cause. Statistical changes due to industry responses to past accidents are controlled for in the modeling.

With this definition, we compiled an original database of as many events as possible over the period 1950 to 2014. To be included in the database, an accident had to be verified by a published source, some of them reported in the peer-reviewed literature, but others coming from press releases, project

documents, public utility commission filings, reports, and newspaper articles. Such an incremental approach to database building has been widely utilized in the peer-reviewed energy studies literature. Hirschberg et al. have constructed the ENergy-related Severe Accidents Database (ENSAD), the most comprehensive database world-wide covering accidents in the energy sector [124]. Flyvbjerg et al. built their own sample of 258 transportation infrastructure projects worth about 90 billion USD [95, 96]. Ansar et al. [17] built their own database of 45 large dams in 65 different countries to assess cost overruns. Also investigating cost overruns, Sovacool et al. [266, 265] compiled a database consisting of 401 electricity projects built between 1936 and 2014 in 57 countries, which constituted 325'515 megawatts (MW) of installed capacity and 8'495 kilometers of transmission lines.

The dataset includes three different measures of accident severity: the industry standard measure, INES; a logarithmic measure of radiation release, NAMS (Nuclear Accident Magnitude Scale) [236]; and the consequences of accidents measured in 2013 US Dollars (USD). The industry standard measure is the discrete seven point INES, defined as [135]: level 0: events without safety significance, level 1: anomaly, level 2: incident, level 3: serious incident, level 4: accident with local consequences, level 5: accident with wider consequences, level 6: serious accident, and level 7: major accident. Levels 1-3 are considered to be "incidents", and levels 4-7 "accidents". The distinction between incidents and accidents is not clear and thus somewhat arbitrary (e.g., see page 152 of [135]). Incidents tend to concern degradation of safety systems, and may extend to include events where radiation was released and people were impacted. However, when the damage and impact to people and the environment becomes large enough, then the event is deemed an "accident". But, there are rules about how many fatalities, or how much radiation release, is necessary to qualify for a specific INES level.

The second measure, NAMS, was proposed as an objective and continuous alternative to INES [236]. The NAMS magnitude is $M = \log_{10}(20R)$ where R is the amount of radiation released in terabecquerels. The constant 20 makes NAMS approximately match INES in terms of its radiation level definitions.

Finally, the main measure used here is the USD consequences/costs due to an event. This cost measure is intended to encompass total economic losses, such as destruction of property, emergency response, environmental remediation, evacuation, lost product, fines, court and insurance claims, etc. In the case where there was a loss of life, we added a lost "value of statistical life" of 6 MM USD per death. The 6 MM USD figure is chosen as a lower bound of the value of statistical life reported

by various US agencies (e.g., the Environmental Protection Agency, Food and Drug Administration, Transportation Department, etc.) [18]. Practically speaking, given that the costs are taken from different sources, it is unlikely that the data truly reflects all relevant costs. While imperfect and controversial, this has the advantage of leading to a single USD metric associated to each event that combines all possible negative effects of the accidents. The costs were standardized by using historical currency exchange rates, and adjusting for inflation to 2013 USD. Adjusting for differing price levels (e.g., because an equivalent event in Switzerland will cost more than one in the Ukraine) was not done because the majority of events belong to countries with a similar price level (US, UK, Japan, and Western Europe), and because the sample is so heavy tailed that adjusting cost within an order of magnitude has little impact on the statistics.

The result of this effort is a unique dataset containing 216 events. Of these events 175 have cost values, 104 have INES values, and 33 have NAMS values (from [236]). The datasets of Sovacool, from the energy studies literature (e.g,. [263], that has been studied [257, 126]) provided a starting point of around 100 events with cost values. For our dataset, Tab. 6.1 lists the 15 most costly events. The data and severity measures are discussed in the following subsection. The dataset has been published online [288], where the public is encouraged to review and recommend additions and modifications with the intention of continually expanding and improving the quality of the data. We believe that this is very important for the proper understanding of risk in nuclear energy. The cost and INES scores are plotted over time in Fig. 6.1. The frequency with which events exceed the threshold of 20MM USD, and the distribution of these excesses will be studied in Sec. 6.4 and Sec. 6.5 respectively. Further, how these quantities have changed in response to the major accidents of Three Mile Island, 1979, and Chernobyl, 1986, will be studied. There are likely to be changes following the major accident at Fukushima in 2011. However, there has been little time to observe improvements, and the cost data is most incomplete in this area: the cost of 18 of the 29 post-Fukushima events contained in the dataset are, as of yet, unknown. Thus, the industry response to Fukushima cannot be quantified in this study.

This dataset dwarfs existing datasets from the academic literature, and is mature enough to justify an analysis. However it is still important to consider the quality of the data, and what methodology is best to handle the limitations. Regarding data completeness, in a rare statistical statement, the IAEA stated, “During the period July 1995-June 1996, the Agency received and disseminated information relating to 73 events - 64 at NPPs [nuclear power plants] and nine at other nuclear fa-

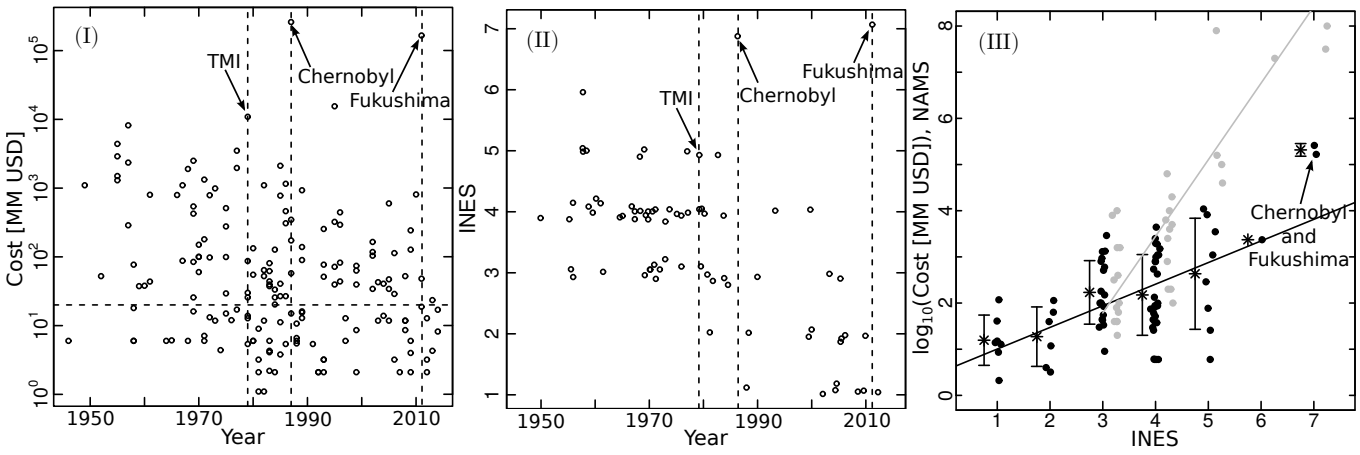

Figure 6.1: Figures concerning the raw data. Panel (I) plots the cost of events over time. The horizontal line is the 20 MM USD cutoff. The vertical lines indicate the occurrence times of TMI, Chernobyl, and Fukushima. Panel (II) plots the INES scores of events over time. The vertical lines again indicate the occurrence times of TMI, Chernobyl, and Fukushima. Panel (III) is a scatterplot of the INES scores and cost of events (black), and the INES scores and NAMS values of events (grey). The star and black vertical lines give the mean and standard errors of the logarithmic costs. The points have been shuffled around their integer INES score for visibility.

cilities. Of those 73 events, 32 were stated to be "below scale" (i.e. safety-relevant but of no safety significance) and three to be "out of scale" (i.e. of no safety relevance). Of the remaining 38 events, three were rated at INES level 3 and eight at level 2 (i.e. as "incidents"), and 27 at level 1 (i.e. as "anomalies")" [132]. On the other hand, in the dataset of this study, only 6 events fall within this period, none of whose INES values are known, rather than the 38. This statistic tells two important things: Firstly, for increasingly small event sizes, our data is increasingly incomplete – the smaller the event, the less likely it is to be discovered, recorded, reported to the regulator, reported in the media, etc. Secondly, our data will remain incomplete for small events until the IAEA publishes historical INES data. The shortage of events of INES level 1, and even 2, prohibits a "near miss" analysis. That is, given that every event develops from having INES score 0, then to 1, and so on, it would be interesting to determine the probability of developing to the next level.

However, the statistics of the largest costs are much more interesting. For instance, the most costly event (Chernobyl) has cost roughly equal to the sum of all other events together, the two most costly events (Chernobyl and Fukushima) have cost roughly five times the sum of all other events together, and the sum of the 53 costs in excess of 100 MM USD is 99.4 percent of the total cost of all 175 cost values. This clearly implies that, if one wants to study the aggregate risk in terms of total

cost, then one simply needs data for the largest events. Thus a typical approach taken in such cases is to study events that are in excess of a threshold, above which the data is thought to be reliable and well modeled. As in [126], we use a threshold of 20MM USD, although results will be similar if a threshold of e.g., 100 MM were used. For cost, there are 101 values above the threshold of 20 MM. Further, of the 41 events with unknown cost, 31 have known INES scores, of which half have level 2 or higher.

Table 6.1: The 15 largest cost events since 1960 are provided with the date, location, cost in MM 2013USD, INES value, and NAMS value. The full dataset is provided online [288]. Unknown values are indicated with a dash.

Date	Location	Cost (MM USD)	INES	NAMS
1986-04-26	Chernobyl, Ukraine	259'336	7	8.0
2011-03-11	Fukushima, Japan	166'089	7	7.5
1995-12-08	Tsuruga, Japan	15'500	-	-
1979-03-28	TMI, Pennsylvania, United States	10'910	5	7.9
1977-01-01	Beloyarsk, USSR	3'500	5	-
1969-10-12	Sellafield, UK	2'500	4	2.3
1985-03-09	Athens, Alabama, United States	2'114	-	-
1977-02-22	Jaslovske Bohunice, Czechoslovakia	1'965	4	-
1968-05-01	Sellafield, UK	1'900	4	4.0
1971-03-19	Sellafield, UK	1'330	3	3.2
1986-04-11	Plymouth, Massachusetts, United States	1'157	-	-
1967-05-01	Chapelcross, UK	1'100	4	-
1982-09-09	Chernobyl, Ukraine	1'100	5	-
1983-08-01	Pickering, Canada	1'000	-	-
1973-09-26	Sellafield, UK	990	4	2.0

6.2.1 Comparing severity measures & critiquing INES

There are many ways to quantify the size of a nuclear accident. Following Chernobyl, several authors proposed to use a monetary value of severity to make events comparable, and use a rate measure normalized by the number of reactor operating years to consider frequency [127, 227, 228]. This is what we have done. Since the IAEA uses INES, it is instructive to compare the two approaches. First, the INES is a discrete scale between 1 (anomaly) and 7 (major accident). Similarly to the Mercalli intensity scale for earthquakes (which has 12 levels from I (not felt) to XII (total destruction)), each level in the INES is intended to roughly correspond to an order of magnitude in severity (in the amount of radiation released).

The INES has been criticized for instance in [236] (and references therein). Common criticisms include that the evaluation of INES values is not objective and may be misused as a public relations (propaganda) tool; moreover, the historical scores are not published, not all events have INES values assigned, no estimate of risk frequency is provided, and so on. Given confusion over the INES scoring of the Fukushima disaster, nuclear experts have stated that the "INES emergency scale is very likely to be revisited" [40]. In [236], the Nuclear Accident Magnitude Scale (NAMS), a logarithmic measure of the radiation release, was proposed as an objective and continuous alternative to INES. This proposition, to go from the INES to the NAMS, is reminiscent of when the geophysics discipline replaced the discrete Mercalli intensity scale by the continuous Richter scale with no upper limit, which is also based on the logarithm of energy radiated by earthquakes. In the earthquake case, the Mercalli scale was invented more than a hundred years ago as an attempt to quantify earthquake sizes in the absence of reliable seismometers. As technology evolved, the cumbersome and subjective Mercalli scale was progressively replaced by the physically based Richter scale. In contrast, the INES scale looks somewhat backward from a technical and instrumental point of view, but was created in 1990 by the International Atomic Energy Agency as an effort to facilitate consistent communication on the safety significance of nuclear and radiological events, while more quantitative measures are available.

Here, we perform a statistical back-test of the accuracy of INES values in relation to costs and NAMS. Indeed INES is not defined in terms of cost, however if INES fails to capture the information that costs do, then the cost measure is important. In Fig. 6.1, we plot both the logarithm of cost, and NAMS versus INES. There is an approximate linear relationship between INES and log cost [intercept parameter at INES= 0 is 0.64 (0.3) and slope 0.43 (0.08) by linear regression]. This is consistent with the concept that each INES increment should correspond to an order of magnitude in severity. However, cost grows approximately exponentially ($10^{0.43} \approx e^1$) rather than in multiples of 10 with each INES level. Further, the upper category (7) clearly contains events too large to be consistent with the linear relationship. For instance, the largest events (Fukushima and Chernobyl) would need to have an INES score of 10.6 to coincide with the fitted line. In addition, the cost of INES level 3 events do not appear to be statistically different from the sizes of INES level 4. Finally, there is considerable uncertainty in the INES scores as shown by the overlapping costs. There is an approximate linear relationship between INES and NAMS [at INES= 3 the intercept is 1.8 (0.9) and slope 1.7 (0.2) by linear regression]. One sees from the points, and from the fact that the slope of the line is greater than 1,

that large radiation release events have been given an INES level that is too small. Furthermore, some INES level 3 events should be INES level 2. This illustrates the presence of significant inconsistency of INES scores in terms of radiation release level definitions.

6.2.2 The current fleet of reactors

One must judge the number of accidents relative to the so called volume of exposure; in this case, the number of reactors in operation. This data was taken from [134] and is plotted in Panel I of Fig. 6.2. The number of reactors in operation grew sharply until 1990 after which it stabilized. The stable level has been supported by growth in Asia, compensating for a decline in Western Europe. A steep drop is observed in the Asian volume where, following Fukushima in 2011, all of Japan's reactors were shut down temporarily until further notice [225]. On the topic of reactors, it is important to note that reactors are somewhat informally classified into generations [7]: Generation I reactors were early prototypes from the 1940s to 1960s. Generation II reactors were developed and built from the 1960s to the 1990s, of which boiling water reactors (BWR) and pressurized water reactors (PWR) are common. Generation III reactors have been developed since the 1990s'. These reactors, such as the advanced BWR and PWR reactors, were improvements upon their Generation II ancestors, replacing safety features that required power with passive ones, and having more efficient fuel consumption. Generation IV reactors concern new technologies, and are still being researched/developed. They have the intention of further improving safety and efficiency, which were deemed as still being inadequate in the Generation III reactors. The vast majority of existing reactors are of Generation II, where most Generation I reactors have been decommissioned, and few Generation III reactors have been constructed. Generation IV reactors are not expected to be deployed commercially until at least 2030 or even 2040.

6.3 Uncertainty quantification of risks in nuclear energy systems

Prior to moving ahead with data analysis and interpretation, reflection on the degree of uncertainty present, and how it is handled in the analysis, is warranted. Following [19] two important questions are (i) if (frequentist or Bayesian) probabilities are attainable, and (ii) if the proposed probabilistic model is accurate/valid. Given a lack of relevant data (e.g., when talking about the future), or difficulty with justifying models, the answer to these questions may be no. If the answer to both

questions is no, then one can be said to be in a state of *deep uncertainty*. Such considerations are relevant to the study of risk in nuclear energy systems, and are discussed below.

Fortunately for this study, the context is clear as historical risks, and the current risk level, are being analyzed. Future risk is only being discussed insofar as the current state remains. Thus, the analysis does not need to deal with uncertain futures. Also, by studying risk at a global level, one avoids epistemic uncertainties associated with the specificities of a given plant, type of accident, or technology. Furthermore, relevant data is available and a simple and somewhat justifiable model used. Thus, here a probabilistic approach is valid, where uncertainties include epistemic model uncertainty, aleatoric statistical uncertainty in the parameter estimates, as well as data uncertainty. The epistemic and aleatoric uncertainties are dealt with by imposing a relatively broad range of parameter estimates (for frequency, severity distribution, and maximum possible severity). Regarding data uncertainty, the data studied here is at a much higher quality level than that of previous studies on nuclear risks. Although the authors are committed to ongoing expansion and refinement of the data, this is a sensible point to provide an analysis. That is, it is unlikely that reasonable modifications of the cost estimates will substantially impact the high-level results provided.

Going beyond this analysis, one can look deeper into nuclear risk. For instance, as regulation requires, PSA is done for each individual plant [129, 130]. This necessitates that the probabilities of initiating events be specified and that the interaction of events be encoded in a fault tree. Further, if one wants to perform level 3 PSA, then the consequences of each event need to be specified [133]. For this task, at a unique power plant, there is little data and thus the huge number of parameters must be specified based on belief/assumption rather than frequentist estimates. Thus, in addition to the aleatoric sampling uncertainty that is captured by simulating from the model, one should also consider the epistemic uncertainty in the model specification and its parameters [88, 87]. Such an approach has been suggested in [208] and a research project considering such methodology in studying the risk of a large loss of coolant is underway [137]. However, standard PSA practice and regulations are not yet at this level of uncertainty quantification. Furthermore, needing to encode all possible events in the fault tree implies that the worst case is limited to the one that the modeler can imagine. Finally, the epistemic uncertainty present in the specification of the (typically deterministic) fault tree, which will practically always be incomplete, is not considered at all. These data and epistemological limitations imply that PSA exhibits deep uncertainty. These issues largely inhibit the ability of standard PSA to

provide an adequate quantification of overall risk.

Nonetheless, PSA is an important and useful exercise. That is, it is a top-down technique that allows for the generation and prioritization of high risk events that may not have been observed, and for common causes to be identified. It is thus instrumental in safety improvement, and risk-informed decision making. The statistical approach is a bottom-up technique, whose specificity (e.g., the risk of a specific reactor technology) is restricted by limited historical data, and whose instances from which one can learn are limited to those that have been observed. As suggested in [171], it is natural to combine top down and bottom up approaches, at least by comparing their results. It is clear that this should be done in nuclear energy systems as well.

Taking a broader perspective, the future risk of nuclear energy should be considered to support decision-making both within the nuclear energy industry, but also within the portfolio of energy source alternatives. The future risk of nuclear energy is deeply uncertain: it depends heavily on developments in reactor and disposal technology, plant build-out and decommissioning scenarios, the emerging risk of cyber-threats and attacks, etc. Furthermore, in making decisions about the holistic plans for future energy systems, the uncertainties of other energy sources also become relevant across multiple criteria. Many risks are reasonably well understood, such as reduced life expectancy, but the evaluation of terrorist threats [123], and the potentially severe environmental impacts of carbon emission are deeply uncertainty [144, 110]. Thus, energy system decisions should be supported by robust multiple criteria decision-making tools with adequate consideration of uncertainty. A probabilistic attempt could be scenario analysis with probability distributions being assigned to the scenarios. Alternatively, non-probabilistic methods may be warranted, as has been done in large-scale engineering systems with multiple diverse stakeholders [145].

6.4 Event frequency

Regarding the frequency of events, we observe $N_t = 0, 1, 2, \ldots$ events each year for the v_t nuclear plants in operation for years $t = 1960, 1961, \ldots, 2014$. The *annual observed frequencies* of accidents per operating facility are $\widehat{\lambda}_t = \frac{N_t}{v_t}$. The observed frequencies are plotted in Panel (II) of Fig. 6.2. The rate of events has decreased, and perhaps stabilized since the end of the eighties. The running rate

estimate,

$$\widehat{\lambda}^{\mathrm{RUN}}_{t_0,t_1} = \frac{\sum_{t=t_0}^{t_1} N_t}{\sum_{t=t_0}^{t_1} v_t}, \tag{6.1}$$

used in [126] is plotted for $t_0 = 1970$ and $t_1 = 1970, 1971, \ldots, 2014$. In the presence of a decreasing rate of events, such a running estimate overestimates the rate of events for recent times. Furthermore, it is a smoothing method where the estimate is taken at the rightmost edge of a constantly growing smoothing window, rather than in the center of a window with fixed width. To avoid this bias and to properly evaluate the trend, we consider another approach. We assume that N_t are independently distributed Pois($\lambda_t v_t$). The Poisson model features no interaction between events, which is sensible as separate nuclear events should occur independently, and is compatible with how we have defined our events in Sec. 6.2. The changing rate of events is accomodated by a log-linear model for the Poisson rate parameter,

$$\lambda_t^{GLM} = \mathrm{E}\left[\frac{N_t}{v_t}\right] = \exp(\beta_0 + \beta_1(t - t_0)), \tag{6.2}$$

for given $t_0 < t$ and parameters β_0, β_1. This is the so-called Generalized Linear Model (GLM) for Poisson Counts [195] and may be estimated by maximum likelihood (using in R:glm). The GLM model was estimated from 1970 until 1986 and from 1987 until 2014, with estimated parameters in Tab. 6.2, and plotted in Panel (II) of Fig. 6.2. The first estimate suggests a significantly decreasing rate, which is in agreement with the decreasing running rate estimate. The second (the approximately flat) GLM indicates that from 1987 onwards the rate has been not significantly different from constant. Clearly the running rate estimate, starting at 1970, is unable to account for this. To further diagnose this difference, in Panel (III) of Fig 6.2, the running estimate (eq. 6.1) is done running both forward and backward from the Chernobyl event of 1987. From here it clear that the rate prior to Chernobyl $\widehat{\lambda}^{\mathrm{RUN}}_{1960,1986} = 0.013$ (0.002), is larger than the rate after $\widehat{\lambda}^{\mathrm{RUN}}_{1987,2014} = 0.0032$ (0.0006). Thus it is apparent that there was a significant reduction in the frequency of events following Chernobyl – likely due to a comprehensive industry response.

It is interesting to note that such a change is not apparent following TMI. There is insufficient data to identify a change following Fukushima. That the data is incomplete implies that our rate estimates are under-estimates. For instance, even within our dataset, there are forty events occurring after Chernobyl whose cost is unknown. Of these forty, thirty-two have INES values, and seventeen of these have INES value of 2 or larger. Based on the known INES and cost values, the median cost of

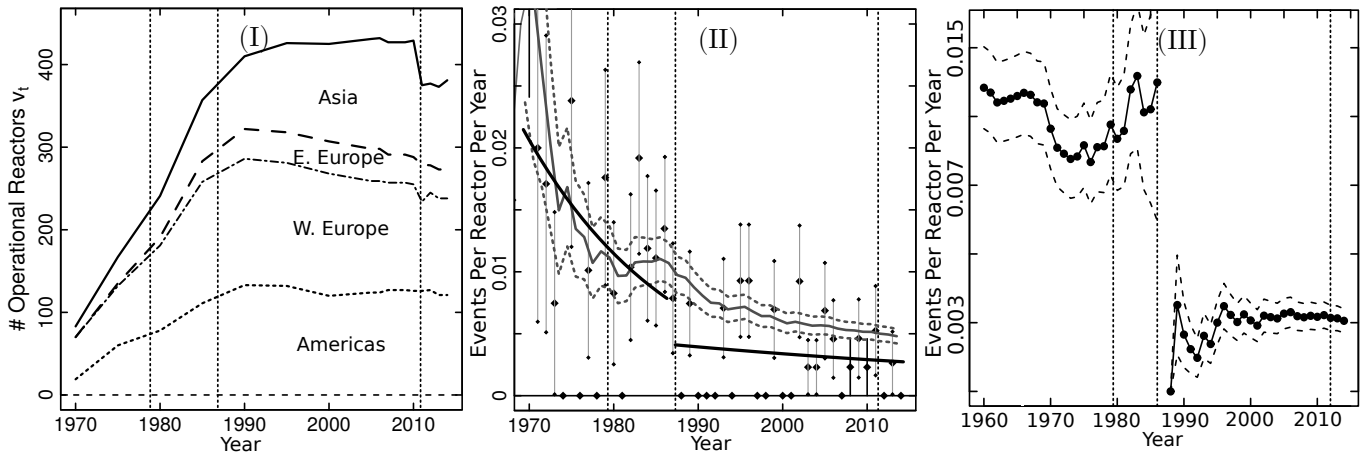

Figure 6.2: Figures concerning the rate of events. Panel (I): The number of operational reactors over time [134] where the bottom layer is the United States, the second layer is Western Europe, the third layer is Eastern Europe, and the top layer is Asia. Panel (II): The annual observed frequencies are given by the solid dots, with Poisson standard errors given by the vertical grey lines ($\mathrm{Var}(\lambda_t) = \frac{\lambda_t}{v_t}$). The solid irregular line bounded by dotted lines is the running rate estimate (eq. 6.1) from 1970 onwards, with standard errors. The solid black lines are the Poisson GLM regressions (eq. 6.2) from 1970 to 1986 and 1987 to 2014, with parameters in Tab. 6.2. Panel (III) provides the running rate estimate (eq. 6.1) running both backwards and forwards from the Chernobyl event in 1987. In all three panels the vertical dotted lines indicate the occurrence of the TMI, Chernobyl, and Fukushima events.

events with INES= 2 is 26 MM USD, i.e., more than half of INES= 2 events exceed the threshold. Thus, based on these statistics, assuming that only nine of the forty unknown events have cost in excess of 20 MM USD is conservative. With such an assumption, the estimate becomes $\widehat{\lambda}^{\mathrm{RUN}*}_{1987,2014} = 0.004$ (0.0006). Taking into account the above, and that the rate of events may actually still be decreasing, we consider a conservative range of current estimates $\widehat{\lambda}_{2014}$ between 0.0025, and 0.0035. This suggests an average of 1 to 1.5 events per year across the current fleet of reactors. Thus, despite having a dataset twice as big as [126], we find a similar rate estimate. Further, provided that the fleet does not undergo any major changes, we expect the rate to remain relatively stable.

As in [126], a significant difference between the frequency of events across regions is found. In Tab. 6.3, one sees that the running estimate of the annual rate varies by as much as a factor of 3 across the regions. This is likely to be more due to a difference in reporting rather than a difference in the true rate of events. This provides further evidence that our rate estimates are underestimates.

Table 6.2: Parameter estimate, standard error, and p-value for GLM estimates of rate (eq. 6.2). The two rows are two estimates for starting times 1970 and 1987. The intercept parameter is given for at the starting time.

Model	β_0	β_1
GLM 1970:1986	$-3.87\ (0.3),\ 10^{-16}$	$-0.06\ (0.03),\ 0.05$
GLM 1987:2014	$-5.53\ (0.3),\ 10^{-16}$	$-0.015\ (0.02),\ 0.45$

Table 6.3: Statistics by region: Number of events (N) and number of reactor years (v) from 1980 through 2014, the rate of events per reactor year, and the Poisson standard error of the rate. Russia is included in Eastern Europe.

Region	N	v	$\widehat{\lambda}^{RUN}_{1980,2014}$	std.
Americas	32	4'378	0.0073	0.001
Western Europe	12	4'813	0.0027	0.001
Eastern Europe	5	2'180	0.0023	0.002
Asia	13	2'548	0.0051	0.001

6.5 Event severity

For the quantitative study of event severity, costs (measured in MM 2013 USD) are considered to be i.i.d random variables $X_i,\ i = 1, 2, \ldots, n$ with an unknown distribution function F. Here, we estimate the cost distribution. A common heavy-tailed model for such applications is the Pareto CDF,

$$F_P(x; u_1) = 1 - (x/u_1)^{-\alpha},\ x > u_1 > 0\ ,\ \alpha > 0\ , \tag{6.3}$$

which may be restricted to a truncated support as,

$$F(x|u_1 \leq X \leq u_2) = \frac{F(x) - F(u_1)}{F(u_2) - F(u_1)},\ 0 < u_1 < u_2\ , \tag{6.4}$$

where u_1 and u_2 are lower and upper truncation points that define the smallest and largest observations allowed under the model. Extending further, truncated distributions may be joined together to model different layers of magnitude,

$$F_{2P}(x|u_1 \leq X) = F_P(x|u_1 \leq X \leq u_2)\Pr\{u_1 \leq X \leq u_2\} + F_P(x|u_2 \leq X)\Pr\{u_2 \leq X\}. \tag{6.5}$$

In Fig. 6.3, the severity measures are plotted according to their empirical complementary cumulative distribution functions (CCDFs). In Panel (I), the sample of costs in excess of 20 MM USD, is split into pre and post TMI periods with 42 and 62 events respectively. The distributions are clearly different. Indeed, the KS test [268], with the null hypothesis that the data for both subsets come from the same model, gives a p-value of 0.015. As can be seen from the lower inset of the first panel of Fig. 6.3, this p-value is much smaller than the p-values obtained for testing other change-times. For instance, there was no apparent change between the pre and post Chernobyl periods. The pre-1979 data, having median cost of 283 MM USD, has a higher central tendency than the post-1979 data, having a median cost of 77 MM USD. However, the post-1979 distribution has a heavier tail, whereas the pre-1979 distribution decays exponentially. It is a rather well-known observation that improved safety and control in complex engineering systems tends to suppress small events, but often at the expense of more and/or larger occasional extreme events [247, 259, 155]. This raises the question of if the observed change of regime belongs to this class, as a result of the improved technology and risk management introduced after TMI.

Thus, we focus on estimating the left-truncated Pareto (eq. 6.3) for the post-1979 data. The estimate $\widehat{\alpha}(u_1)$ fluctuates in the range of 0.5-0.6 for lower threshold $20 < u_1 < 1000$ MM USD, indicating that the data is consistent with the model. For $u_1 < 20$, the estimate of α is smaller, as is typical for datasets where small events are under-reported. In [257], the estimated value was larger ($\alpha = 0.7$), while [126] also found values between 0.6 and 0.8. With our more complete dataset, the smaller value α is qualitatively consistent with previous studies, but further emphasizes the extremely heavy tailed model ($\alpha \leq 1$) where the mean value is mathematically infinite. In practice, this simply means that the largest event in a given catalog accounts for a major fraction ($\sim 1 - \alpha$) of the total dollar cost of the whole [243]. That is, the extremes dominate the total cost.

6.6 Runaway disasters as "dragon king" outliers

In a complex system with safety features/barriers, once an event surpasses a threshold, it can become uncontrollable, and develop into a "runaway disaster" – causing disproportionately more damage than other events. This is the type of phenomenon that is considered here. In Panel (I) of Fig. 6.3 one can see that (at least) the two most costly events since TMI (Chernobyl and Fukushima) lay above the estimated Pareto CCDF, and that Chernobyl, TMI, and Fukushima form a cluster of outliers in

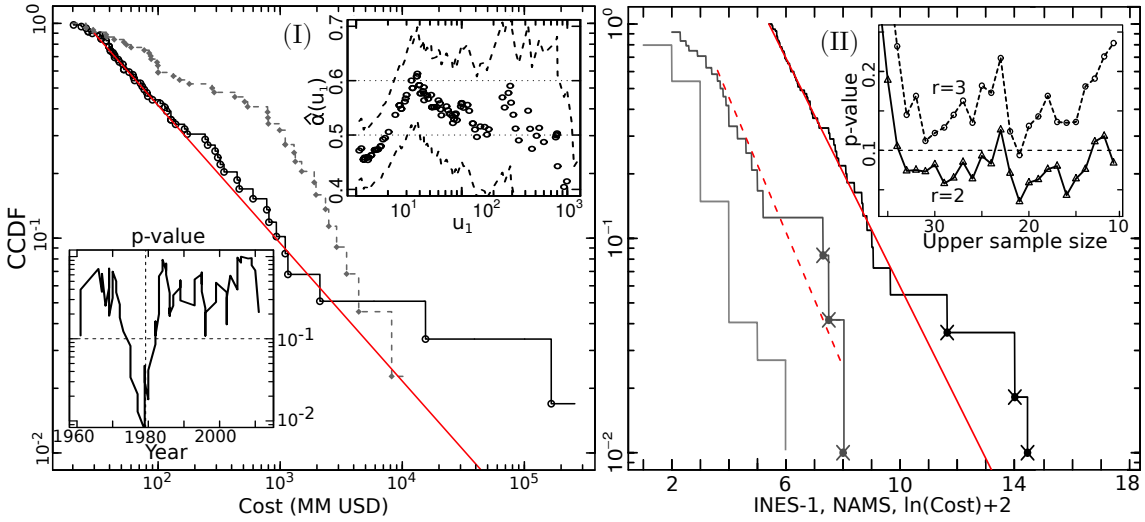

Figure 6.3: Figures concerning the severity of events. Panel (I): The main frame plots the cost of events for the pre and post-TMI periods according to their CCDFs, in grey and black respectively. The lower inset figure shows the p-value of a segmentation test of the cost data, identifying TMI (1979) as the change-point in the cost distribution. The upper inset figure shows the estimated parameter α (with standard deviation) of a Pareto distribution (eq. 6.3), for the post-TMI cost data, for a range of lower thresholds (u_1). The fit for $u_1 = 30$ (MM USD) is given by the red solid line in the main frame. Panel (II): In the main frame, from left to right, are the CCDF of INES scores above 2 (shifted left by 1), the CCDF of NAMS scores above 2, and the CCDF of the natural logarithm of post 1980 costs (shifted right by 2). For the center and right CCDFs, the dots with x marks indicate suspected outliers/dragon-kings. The dashed and solid red lines are exponential fits to the CCDFs. The inset figure provides the p-value for the outlier test for the upper sample above a growing threshold. The upper curve is for $r = 3$ outliers, and the lower curve for $r = 2$ outliers.

the sample of NAMS data. The NAMS data cannot be split into pre and post TMI samples due to insufficient data. The NAMS distribution is well described by the Exponential distribution for values between 2 until 5, as reported in [236]. That is, for NAMS values above $u_1 = 3.5$, the estimate is $\widehat{\alpha}_{NAMS} = 0.72$ (0.3) with sample size $n = 15$, where the three suspected outliers were censored to avoid biasing the estimate [69]. Since NAMS is a logarithmic measure of radiation, this corresponds to a Pareto distribution for the radiation released, which is valid over 3 decades.

We relate this concept of runaway disasters to the concept of dragon-kings. The term dragon-king has been introduced to refer to such situations where extreme events appear that do not belong to the same distribution as their smaller siblings [247, 259]. Dragon-king (DK) is a double metaphor for an event that is both extremely large in size or impact (a "king") and born of unique origins (a "dragon") relative to other events from the same system. For instance, a DK can be generated by a transient

positive feedback mechanism. The existence of unique dynamics of DK events gives hope that they may be "to some extent" predictable. In this sense, they are fundamentally different from the a priori unpredictable "black swans" [271, 20]. Statistically speaking, given that extreme events tend to follow heavy-tailed distributions, DK can be specifically characterized as outlying large extremes within the population of extreme events. That is, DK live beyond power law tails, whereas black swans are often thought of as being generated by (unanticipated) heavy power law tails.

Let us now make the above observations more rigorous by testing the apparent DK points as statistical outliers. There are many tests available to determine if large observations are significantly outlying relative to the Exponential (or Pareto) distribution [23, 270, 201] with a recent survey in [290]. A suitable approach to assess the NAMS outliers is by estimating a mixture of an Exponential and a Normal density,

$$f_{\text{NAMS}}(x|x > 3.5) = \pi\alpha\exp\{-\alpha x\} + (1-\pi)\phi(x;\mu,\sigma) \ , \ \alpha,\sigma > 0 \ , \tag{6.6}$$

where the Gaussian density $\phi(x;\mu,\sigma)$ provides the outlier regime, and $0 \leq \pi \leq 1$ is a weight. The test is done for the 14 points in excess of 3.5, where the Exponential tail is valid. The Maximum Likelihood estimation of this model (eq. 6.6) is done using an Expectation Maximization algorithm [210, 32]. The estimates of this (alternative) model are $(\widehat{\pi} = 0.8, \widehat{\alpha} = 0.78, \widehat{\mu} = 7.8, \widehat{\sigma} = 0.21)$. We also consider a null model with no DK regime ($\pi = 1$). For this the MLE is $\widehat{\alpha} = 0.66$. The alternative model has a significantly superior log-Likelihood (the p-value of the likelihood ratio test [293] is 0.04). Thus there is a statistically significant DK regime relative to the Exponential, with $(1-\hat{\pi})\cdot 14 \approx 3$ outliers expected.

That the amount of cost is related to the amount of radiation released suggests testing for a DK regime in cost. Not every runaway radiation release disaster produces commensurate financial damage (see Three Mile Island in Tab. 6.1). But, given that the majority of nuclear power installations have surrounding population densities higher than Fukushima [4], the DK regime in radiation should amplify cost tail risks. From Panel (II) of Fig. 6.3 as many as the three largest points could be outlying. For this we consider the *sum-robust-sum* (SRS) test statistic,

$$T_r^{SRS} = \frac{\sum_{i=1}^{r} x_{(i)}}{\sum_{i=r+1}^{n} x_{(i)}}, \quad m \geq 1 \ , \tag{6.7}$$

for the ordered sample $x_{(1)} > x_{(2)} > ... > x_{(n)}$, which compares the sum of the outliers to the sum

of the non-outliers [290]. This test was performed for $r = 2$ and $r = 3$ outliers for a range of upper samples – i.e., the sample in excess of a growing lower threshold. For $r = 2$, the p-value fluctuates between 0.05 and 0.1 for samples ranging from the ten to the forty largest points. For $r = 3$, the test fluctuates between 0.1 and 0.2. Thus, there is evidence that the two largest events are indeed outliers, both in terms of radiation and cost.

Given the suggestive evidence that the extreme tail of cost is heavier than the rest of the tail, perhaps due to a runaway disaster effect, it is important to include this in the risk modeling. For simplicity, we continue with the Pareto model. The MLE for the top 5 points is $\widehat{\alpha}(u_1 = 1100) = 0.36$ (0.15). To pursue a pleasant but non-rigorous argument, this appears to be consistent with the run-away effect that propels NAMS values from $M \approx 5$ to $M \approx 8$. That is, transforming back from log scale, this same effect on the Pareto model would transform the parameter α to $\alpha(u_1 = 1100) \approx \frac{5}{8} \times 0.61 = 0.375$.

6.7 Modelling aggregate annual damage

We now characterize the annual risk of nuclear events measured by cost, with three quantities: quantiles, return periods, and expected values. These characterizations are relevant for the current state of the operating nuclear fleet – excluding any potential improvements following Fukushima – and do not consider scenarios for the transition to more advanced reactor technologies, such as Generation III and beyond.

6.7.1 Quantiles and return periods

Having estimated the rate of events in Sec. 6.4, the severity distribution in Sec. 6.5, and identified the DK regime in Sec. 6.6, we here combine these models in a Compound Poisson Process (CPP) (for references, e.g., [296, 182]) to model the annual total cost in MM USD,

$$Y_t = \sum_{i=1}^{N_t} X_{i,t} \ \sim \text{CompPois}(v_t\lambda_t, F) \ , \tag{6.8}$$

where, for each year $t = 1980, 1981, \ldots, 2014$, there are a random number of events N_t, modeled by a Poisson process with annual rate $v_t\lambda_t$, and each event has a random size (in MM USD) $X_{i,t} \overset{i.i.d.}{\sim} F,\ F(20) = 0,\ i = 1, \ldots, N_t$, where the condition $F(20) = 0$ ensures that we consider only events with

costs larger than 20 MM USD.

There are a range of statistically valid parameter estimates that should be considered. From Sec. 6.4, rate estimates ranging between 0.025 and 0.035 were suggested as conservative underestimates. From Sec. 6.5, the distribution of cost in excess of 20 MM was found to be well described by a Pareto distribution with parameter between 0.5 and 0.6. Further, in Sec. 6.6 it was found that the largest cost values are significantly larger than what would be expected under the Pareto model. An attempt to account for this was made by including a heavier tail (a DK regime) with $\alpha \approx 0.4$ for the top ten percent of the mass (with lower threshold $u_1 = 1100$ MM USD).

First we characterize the risk level with *return periods*, defined within the CPP model by considering,

$$\Pr\left[\{\#\text{ events with size} \geq x_{(j)} \text{ in } \tau \text{ years} > 0\}\right] = 1 - \exp\left[-\lambda v \tau \Pr\{X \geq x_{(j)}\}\right] \ , \tag{6.9}$$

which is the probability of observing at least one event, at least as large as some size (e.g., given by an order statistic $x_{(j)}$), in a given time period τ. One sets equation (eq. 6.9) to a given probability p and solves for the return period $\tau_j(p)$ of the j^th largest event. Setting $p = 1 - e^{-1}$, one obtains the standard return period $\tau_j(p) = \frac{1}{\lambda v \Pr\{X \geq x_{(j)}\}}$. In Tab. 6.4, median and quartile estimates of $p = 0.5$ return periods estimates are given for combinations of the above range of parameter values. For each given set of parameter values, 100'000 samples of the data were simulated, parameters re-estimated on these samples, and the return period computed. The median and quartiles were taken over these 100'000 return period estimates. In the lowest risk model ($\lambda = 0.00275$, and $\alpha = 0.6$), the median $p = 0.5$ return periods for TMI, and Fukushima are 17, and 154 years respectively. In the most conservative case ($\lambda = 0.0035$, and $\alpha = 0.4$ in the extreme tail), these values are 12, and 62. It is clear that including the DK effect does not inordinately amplify the risk. For instance, the return periods with high frequency risk ($\lambda = 0.0034$) and low risk severity ($\alpha = 0.6$) are similar to those of the low frequency frequency risk ($\lambda = 0.00275$) and high severity risk ($\alpha = 0.4$ in the extreme tail).

Next we provide quantiles. For F, we take the estimated Pareto cost distribution (eq. 6.3 with $u_1 = 20$ and $\widehat{\alpha} = 0.55$). We also consider F_{DK}, which is a two layer model (eq. 6.5) where the upper layer is for the DK regime. The first layer, from $u_1 = 20$ to $u_2 =$1'100, is Pareto with $\widehat{\alpha_1} = 0.55$ (0.15) estimated by MLE. The second layer, from $u_2 =$1'100 onwards, is also Pareto with heavier tail $\widehat{\alpha_2} = 0.4$. Given v_t, λ_t, and F, we can calculate the "aggregate" distribution G for annual cost Y_t. We do this for

Table 6.4: Median and quartile $p = 0.5$ return periods (eq. 6.9) for Fukushima ($x_{(2)}$), and TMI ($x_{(4)}$). for different rate parameters λ and parameters α for the Pareto distribution above lower threshold u. The last column corresponds to the "dragon king" regime. The median and quartiles are computed over 100,000 estimates of the parameter values computed on data simulated given the parameter values provided in the table.

λ	$x_{(j)}$	$\alpha = 0.6,\ u = 20$	$\alpha = 0.55,\ u = 20$	$\alpha = 0.4,\ u = 1100$
0.00275	$x_{(2)}$	$(97, 154, 259)$	$(64, 98, 157)$	$(37, 82, 274)$
0.00275	$x_{(4)}$	$(13, 17, 24)$	$(10, 13, 18)$	$(10, 15, 30)$
0.0035	$x_{(2)}$	$(80, 128, 214)$	$(50, 80, 130)$	$(31, 66, 198)$
0.0035	$x_{(4)}$	$(10, 13, 19)$	$(8, 10, 14)$	$(8, 12, 22)$

the year 2014 with the Panjer algorithm by Monte-Carlo [296, 182]. Quantiles of the estimated G are in Tab. 6.5. The 0.99 quantile is highly sensitive to the choice of λ and distribution F: For very low rate $\lambda = 0.002$, and without considering the DK effect, the 0.99 quantile is 54'320 (MM USD) which is five times the cost of TMI. For $\lambda = 0.0025$, we obtain a similar estimate to [126], who obtained 81'000 (MM USD). Considering $\lambda = 0.003$, with the DK effect, this quantile is 331'610 (MM USD), which is double the estimated cost of Fukushima.

Table 6.5: The estimated 0.95 and 0.99 quantiles, as well as the probability of the annual cost exceeding the cost of Fukushima $x_{(2)} = 166,089$ MM USD, are given for the aggregate distribution G. The Pareto model is with $u_1 = 20,\ \alpha = 0.55$, and the Pareto DK model is with $u_1 = 20\ u_2 = 1100,\ \alpha_1 = 0.55,\ \alpha_2 = 0.4$. The volume (number of active nuclear plants) is taken to be $v_{2014} = 388$. The quantiles are given in MM 2013USD.

Model	λ	$q_{0.95}$	$q_{0.99}$	$\Pr\{Y_t \geq x_{(2)}\}$
Pareto	0.002	2'950	54'320	0.0054
	0.0025	4'440	82'440	0.0068
	0.003	6'200	115'780	0.0082
Pareto DK	0.002	2'180	120'730	0.0088
	0.0025	3'720	220'510	0.011
	0.003	5'880	331'610	0.013

6.7.2 Expected annual damage

So far we have considered models without a limiting cost, in which the mean cost is mathematically infinite, since our various estimations of the Pareto exponent α all converge to values less

than 1 [243]. Of course, the Earth itself is finite, thus there is an upper cut-off, u_2, to the maximum possible cost. But this upper cut-off could be exceedingly large, and there – as of yet – no evidence of a maximum being reached thus far (i.e., no accumulation of observations at an upper limit in Fig. 6.3). Think for instance of the real-estate value of New York City, USA or Zürich, Switzerland, both of which are rather close to a nuclear plant, and would become inhabitable in a worst case scenario. Here, we would be speaking of up to tens of trillions of dollars of financial losses, not to speak of human ones. Thus, insurance and re-insurance companies introduce a maximum loss for their liabilities, which for them works as if there is a genuine upper cut-off: u_2. Everything above such a cut-off is then the responsibility of the government(s) and society; for the truly extreme catastrophes, only the state can be the insurer of last resort.

It is useful to put hard numbers behind these considerations by using scenarios. For the CPP (eq. 6.8), the mean and variance of the annual cost are

$$E[Y_t] = \lambda_t v_t E[X], \quad Var(Y_t) = \lambda_t v_t E[X^2] \ . \tag{6.10}$$

Given lower and upper truncations, u_1 and u_2, the first two moments for the Pareto are,

$$E[X] = \frac{\alpha}{\alpha - 1}\left[\frac{u_1^{1-\alpha} - u_2^{1-\alpha}}{u_1^{-\alpha} - u_2^{-\alpha}}\right] \ , \ E[X^2] = \frac{\alpha}{\alpha - 2}\left[\frac{u_1^{2-\alpha} - u_2^{2-\alpha}}{u_1^{-\alpha} - u_2^{-\alpha}}\right] \ . \tag{6.11}$$

Thus the mean grows in proportion to $u_2^{1-\alpha}$ (and the variance faster as $u_2^{2-\alpha}$). In Tab. 6.6, we compute these moments of the costs X when the maximum value, u_2, is equal to the present estimate of the cost of Fukushima, ten times greater, and one hundred times greater. Since the expected annual number of events $\widehat{\lambda}_{2014} v_{2014}$ is approximately 1, these values provide a rough estimation of the mean and standard deviation of annual cost in 2014 (eq. 6.10).

Table 6.6: The first moment and the square root of the second moment of X are given by the first and second value respectively. The Pareto model is with $u_1 = 20$, $\alpha = 0.55$ and three values for the maximum value u_2. The Pareto DK model is with $u_1 = 20$ $u_2 = 1100$, $\alpha_1 = 0.55$, $\alpha_2 = 0.4$ and three values for the maximum value u_3. The maximum values are 1, 10, and 100 times the cost of Fukushima, $x_{(2)}$ =166'089 MM USD. All units of costs are in MM USD.

Model	$x_{(2)}$	$10 \times x_{(2)}$	$100 \times x_{(2)}$
Pareto	1'513, 8,253	5'367, 54'590	20'488, 349'736
Pareto DK	1'404, 8,466	3'982, 45'267	11'250, 240'810

If we accept that the Fukushima or Chernobyl events represent roughly the largest possible cost then (see Tab. 6.6) the mean annual cost is approximately 1.5 Billion USD with a standard error of 8 Billion USD. This brackets the construction cost of a large nuclear plant, suggesting that about one full equivalent nuclear power plant value could be lost each year, on average. However the heavy tailed severity implies that most years, there is little cost, and once in a while an extreme hits, driving the total cost up considerably. If we assume that the largest typical possible cost is about 10 times that of the estimated cost of Fukushima, then the average annual cost is about 5.5 Billion USD with a very large dispersion of 55 Billion USD. Indeed, the outlook is even more dire for larger possible upper-cutoffs. Such numbers do not appear to be taken into account in standard calculations on the economics of nuclear power (see for instance [294])). To be fair, we should also note that the long-term effect on, say, lung cancer risks and other particle pollution induced deaths, are not taken into account in evaluating the cost-benefits of alternative sources of energy such as coal.

6.8 Discussion & policy conclusions

Our study makes important conclusions about the risks of nuclear power. Regarding event frequency, we have found that the rate of incidents and accidents per civil nuclear installations decreased from the 1970s until the present time. Along the way, there was a significant drop in the rate of events after Chernobyl (April, 1986). Since then, the rate has been roughly stable, implying a rate between 0.0025 to 0.0035 events per plant per year in 2015. It is worth noting that the decrease in risk due to the reduced accident frequency per reactor from the 1960s onwards has been somewhat offset by an increasing number of reactors in operation.

Regarding event severity, we found that the distribution of cost underwent a significant regime change shortly after the Three Mile Island major accident. Moderate cost events were suppressed but extreme ones became more frequent to the extent that the costs are now well described by the extremely heavy tailed Pareto distribution with parameter $\alpha \approx 0.55$. We noted in the introduction that the Three Mile Island accident in 1979 led to plant specific full-scope control room simulators, plant specific PSA models for finding and eliminating risks and new sets of emergency operating instructions. The change of regime that we document here may be the concrete embodiment of these changes catalyzed by the TMI accident. We also identify statistically significant runaway disaster ("dragon-king") regimes in both NAMS and cost, suggesting that extreme events are amplified to values even larger than those

explained under the Pareto distribution with $\alpha \approx 0.55$.

In view of the extreme risks, the need for better bonding and liability instruments associated with nuclear accident and incident property damage becomes clear. For instance, under the conservative assumption that the cost from Fukushima is the maximum possible, annual accident costs are on par with the construction costs of a single nuclear plant, with the expected annual cost being 1.5 billion USD with a standard deviation of 8 billion USD. If we do not limit the maximum possible cost, then the expected cost under the estimated Pareto model is mathematically infinite. Nuclear reactors are thus assets that can become liabilities in a matter of hours, and it is usually taxpayers, or society at large, that "pays" for these accidents rather than nuclear operators or even electricity consumers. This split of incentives improperly aligns those most responsible for an accident (the principals) from those suffering the cost of nuclear accidents (the agents). One policy suggestion is that we start holding plant operators liable for accident costs through an environmental or accident bonding system [264], which should work together with an appropriate economic model to incentivize the operators.

Third, looking to the future, our analysis suggests that nuclear power has inherent safety risks that will likely recur. With the current model – which does not quantify improvements from the industry response to Fukushima – in terms of costs, there is a 50% chance that (i) a Fukushima event (or larger) occurs in 62 years, and (ii) a TMI event (or larger) occurs in 15 years. Further, smaller but still expensive ($\geq$ 20 MM 2013 USD) incidents will occur with a frequency of about one per year, under the assumption of a roughly constant fleet of nuclear plants. To curb these risks of future events would require sweeping changes to the industry, as perhaps triggered by Fukushima, which include refinements to reactor operator training, human factors engineering, radiation protection, and many other areas of nuclear power plant operations. To be effective, any changes need to minimize the risk of extreme "dragon-king" disasters. Unfortunately, given the shortage of data, it is too early to judge if the risk of events has significantly improved post Fukushima. We can only raise attention to the fact that similar sweeping regime changes after both Chernobyl (leading to a decrease in frequency) and Three Mile Island (leading to a suppression of moderate events) failed to mitigate the very heavy tailed distribution of costs documented here.

A separate conclusion of our article concerns the nature of data about nuclear incidents and accidents. We found that the INES scale of the IAEA is highly inconsistent, and the scores provided by the IAEA incomplete. For instance, only 50 percent of the events in our database have INES

scores. Further, for the costs to be consistent with the INES scores, the Fukushima disaster would need to be between an INES level of 10 and 11, rather than the maximum level of 7. The INES scale was compared to the antiquated Mercalli scale for earthquake magnitudes, which was replaced by the continuous physically-based Richter scale. Clearly an objective continuous scale such as the NAMS would be superior to the INES. However, while using INES, scores should be made available for all accidents. When such a framework is established, and data on incidents and accidents made more rigorous, and transparent, accident risks can be better understood, and perhaps even minimised through positive learning.

Finally, our study opens a number of avenues for future research. Our results have been obtained for the current fleet, dominated in large part by Generation II reactors. Future research directions could be to investigate how much of the specific risks for each reactor type or design can be inferred from statistical analysis, with the goal of identifying which of the reactors are the safest. In addition to the role of technology, another natural extension would be to correlate accidents to the type of market or form of regulatory governance, restructured versus monopoly/state run, or limited liability versus no limited liability.

Speaking in comparative terms, our focus on the risks of civil nuclear power plants might give the impression that this technology is riskier than other competing technologies, such as coal or wind energy. However, due to the more diluted nature of the costs, and the quasi-hysteric focus on nuclear risks following the Fukushima disaster, an insidious villain may be hidden: it has been estimated that fine particle air pollution causes about seven million premature deaths globally each year, including more than one million in China, and about 60'000 in Europe. Considering the value of life to be in the millions (10^6), these deaths alone account for a cost on the order of a trillion USD (10^{12}), not taking into consideration the billions also being required for health care. Coal, whose global use has soared by 50% from 2000 to 2010, is the leading source of fine particles, which are embedded in the lungs, causing cancers. Between 2010 and 2012, European coal consumption jumped 5%, or 50 million tons. Thus, performing a rigorous empirically based comparative analysis of the risks of nuclear versus other forms of energy providers is absolutely essential to avoid falling in the traps of media hypes and availability biases, in the goal of a better steering of our societies. Furthermore, such an analysis should, on one hand, take into consideration the costs of the disposal of nuclear wastes, while on the other hand recognizing that Humankind is confronted with a "nuclear stewardship curse", whereby

existing nuclear by-products/waste need to be securely managed over immense time scales [248].

6.9 Future Work

Looking forward, we are in the process of expanding the dataset, and constructing better cost estimates. This involves searching for new events, comparing multiple cost estimates for each event, where is estimates are available, and potentially constructing estimates for other events based on comparison to other events. For future work, interesting directions are: 1) Extending the statistical risk analysis of nuclear power in to consider different reactor designs/generations, and linking these analyses to specific PRA estimates. This will make our estimates from more relevant to the recent generations of nuclear power plants and position us to consider future risk scenarios. 2) Publishing the nuclear power incident/accident data from on our ETH website, in a supervised open format where the public can contribute. This will provide public visibility and scientific robustness into the study of the risk of nuclear accidents. This is enormously important. 3) Comparative risk analysis of relevant energy sources (including nuclear, fossil-based, hydro, solar, wind, etc.) based on cost. Essentially, the cost values will include fatalities, rather than considering fatalities separately as has been done in the past. This allows for better comparison of sources like hydro and nuclear, which have historically been high and low fatality risks, respectively. This analysis would bring together the ENSAD data, our nuclear events data [124], as well as data on solar, wind, etc. accidents from Prof. Sovacool. 4) Construct and evaluate the risk of scenarios for future nuclear power plant build-out and decommissioning. This will be based on a combination of statistical risk estimates and PSA estimates for reactors of the new generation for which accident data is not yet available.

Chapter 7

Cyber risk

This chapter is based on [289].

7.1 Introduction

Both the Earth and humanity are often hit by extreme disasters characterised by their high severity and unpredictability [200]. Natural disasters – such as floods, earthquakes, and hurricanes – and man-made disasters – such as financial crashes [148, 242], nuclear power plant meltdowns [291, 257], military nuclear accidents [223], the 2003 space shuttle explosion [161] and many other extreme industrial disasters [199, 49] – belong to this class. A recent entrant into this class are cyber attacks [3], in which intellectual property can be ex-filtrated, and operations of information systems – or even critical physical infrastructures [57, 221] – disrupted. These cyber attacks can be perpetrated by cyber criminals, or even by state actors as acts of espionage or cyber warfare. Here we focus on personal data breaches, as a subset of cyber risks, where large amounts of personal information (i.e., name, social security number, address, email, date of birth, credit card numbers, user-names and passwords, etc.) are ex-filtrated from organisations, typically for use in identity fraud.

While appearing minor in view of other cyber risks, personal data breaches heavily impact both consumers and organisations [113, 202]. For instance, the average financial loss due to the theft of a single piece of private data is estimated to be 213 USD [202]. Thus it is not surprising that data breaches can result in immediate negative impacts in the stock value of traded companies [43, 100, 8, 101]. Further, it is estimated that data breaches (including intellectual property) cost large companies more than 1 trillion US Dollars in 2008 [180]. In fact, the potential for major damage is

so severe that the statistical properties of data breaches have been found to be similar to those of the most extreme disasters mentioned above [174]. Acknowledging this, both governments, insurance companies, and organisations have started ranking cyber risk as one of the largest risks that they face (e.g., [89, 206, 284, 6, 14]).

Understanding the risk of disasters is essential for the proper design of infrastructures, emergency response planning, and for the construction of sound insurance policies [79, 82]. Modern approaches to this involve the systematic collection of high quality data and subsequent statistical risk analysis [79, 200].

Early work on personal data breach risks [174] has demonstrated the Pareto (Power Law) distribution function (df) of large breaches, which, having parameter $\alpha \approx 0.7$, is so *extremely heavy tailed* that the largest observation is expected to be $(1-\alpha)/\alpha \approx 0.43$ times as large as the sum of the rest of the data [244]. For this reason, the mean (and higher moments) are infinite, and thus data breaches constitute an extreme statistical risk. Motivated by the major implications of such a dire characterisation, the fact that, in the past six years since [174] was done, much more data has become available, and given the dynamic nature of the cyber space, we both significantly update and extend this analysis. With a current and much larger dataset, we confirm the heavier tailed breach df, and stable frequency since 2007. Going beyond this, we find that the breach df is in fact even heavier tailed than expected, and has a finite maximum. Further, we explicitly evaluate the connection between organisation size and breach risk.

Specifically, we characterise the risk of large data breaches of personal data occurring at organisations. Damage is measured by the number of individual information items (ids) ex-filtrated. The recording of such data started in the 2000s, by an early online community, scouring media and other online reports [196], and has since become more mature with a variety of communities and organisations taking on the job [204, 282]. Within the United States, this task has been aided by Freedom of Information (FOI) requests, and the onset of legal disclosure obligations faced by organisations having suffered a data breach (Data Breach Notification Act). However, perhaps due to the relatively short history of cyber risks, these databases are not without their weaknesses, including incompleteness, lack of standardisation, in-availability, and so on. Here, we have joined together the two largest databases [196, 204], and filtered for events that occurred at an organisation (both public and private, businesses, universities, hospitals, etc.), within any country. This yielded 6,422 data breach events, each having

in excess of 10 ids breached, between January 1st, 2000 and April 16, 2015, providing a solid basis for statistical analysis that allows us to estimate both the frequency and severity of events. For this modeling, we focus exclusively on large breaches (having in excess of fifty thousand ids). The severity is represented by a Pareto df with extensions allowing for the statistical hypothesis testing of (i) if there is a maximum breach size, (ii) if this maximum is growing, and (iii) if breaches tend to be getting larger over time. These models are tested against one-another and individually verified by rigorous diagnostics. Acknowledging the fact that breached data accumulates in the hands of cyber criminals, we also model the cumulative sum of large breaches over time. This brings together models for both frequency and severity, and provides a gloomy forecast for risk of data breaches, and thus the state of privacy, in the following five years.

Further, extreme statistical properties of disasters may often be related to a complex hierarchical underlying structure, in which cascades of failures develop into a broad df of sizes. For instance, the Gutenberg-Richter (heavy-tailed power law) frequency-magnitude law for earthquakes is thought to originate from the hierarchy of fault scales forming complex fault networks [226]. The proximate trigger of the 2008 financial meltdown can be attributed to a collapse of the hierarchical inter-bank network [237], when overnight loans backed by financial asset collaterals froze [143, 230]. Financial bubbles develop from the interaction of many different agents at different scales– computers, individuals, investment floors, and firms to currency blocks – at different time scales [252].

Similarly, the Internet exhibits a socio-technological complexity that spans all levels of interactions. With fast-evolving hardware and software structures, and coupled with heterogeneous and simultaneous interactions of millions of users, there are many potential points of failure. More specifically, data breach risk is related to underlying factors such as the *attack surface*, which is the number of points where an attacker can extract data, as well as the volume and value of information assets that an organisation is guarding. As a single proxy risk factor for these variables, we consider organisation size, as defined by market capitalisation. With this measure, we unearth how both the frequency and magnitude of breach risk scale with organisation size – thus quantifying the effective *risk surface* of an organisation.

7.2 Data, Results & Methods

The risk of data breaches is considered where each breach has an event time, being the date at which it was reported (to a governmental body or media outlet), and a size, being the number of pieces of personal information breached.

The risk is decomposed into frequency and severity components, and these two components are studied separately. Only events above a threshold size are considered. This is sensible for (at least) two reasons: Firstly, large breaches dominate the total number of breached ids. For instance, although events with size above 5×10^4 only represent less than 10 percent of events, the total number of ids' breached across these large events is above 99 percent of the total across all events (Tab. 7.1). Thus, we want good data for reliable statistics about large breaches. This relates to the second reason: large breach events are more visible, and thus the data are more complete and reliable in this range. Including smaller events in the data set worsens data quality and may bias statistics. We choose the breach size threshold to be 5×10^4, and breaches above this threshold will be referred to as *large breaches*. To further select reliable data, we study events occurring from January 1^{st}, 2007 until April 15^{th}, 2015, being the most recent date available at the time of analysis. Data is available prior to 2007, however the number of events are relatively few, and the statistics less consistent.

It is important to note that almost 40 percent of events have an unknown size (Tab. 7.1). If we exclude these events, and if these events with unknown size tend to be small (i.e., falling beneath the threshold of 5×10^4), then there is no problem. This may well be the case. If they also tend to be large, then we underestimate the frequency. If they tend to be large, and follow a different distribution than the rest of the sample, then this will bias the estimate of the distribution. However, given a lack of covariates to identify if there is a way to predict if an event will have an unknown size, we can neither evaluate the effect of this, neither perform any meaningful imputation (i.e., estimate the missing values in a way that is consistent with the result of the data). That is, any imputation will only sample these events from the distribution identified by the observed events, and thus have no impact on the distribution. Further, this will require an assumption about what proportion of the events with unknown size will be above the threshold. Thus, we simply omit the events with unknown size, providing an optimistic quantification of the risk.

To enable rigorous statistical modeling, we introduce some notation. We consider the *large breach sizes* $\{x_i,\ i = 1, 2, \ldots, n = 619\}$, having limited size $5 \times 10^4 < x_i \leq \nu$, with given lower threshold

5×10^4, and parameter ν gives the unknown maximum breach size. The data x_i are ordered by their event time $0 < t_1 < t_2 < \cdots < t_n$, where the clock starts at $t = 0$, being January 1^{st}, 2007, and one unit of t is a year.

Category	n	Total Breach	Monthly Rate	Annual Rate	GLM
US	6142	1.189×10^9	62.6 (13.1)	751.0 (110.9)	-
US$> 5 \times 10^4$	407	1.174×10^9	4.25 (1.82)	49.5 (6)	4.59 (0.5); -0.08 (0.1), 0.43
Non-US	1978	0.794×10^9	24.6 (18.0)	296.0 (199.2)	-
Non-US$> 5 \times 10^4$	186	0.788×10^9	2.48 (1.65)	26.0 (11.4)	1.64 (0.4); 0.19 (0.1), 0.02
All	8574	1.983×10^9	87.2 (27.1)	1046.5 (292.1)	-
All$> 5 \times 10^4$	619	1.962×10^9	6.29 (2.65)	75.5 (10.38)	5.44 (0.6); 0.18 (0.12), 0.13

Table 7.1: The number of events (n), total number of breached ids (Total Breach), average monthly count (Monthly Rate) and standard deviation of monthly counts, average annual count (Annual Rate) and standard deviation of annual counts, and GLM summary. The generalised linear model (GLM) summary provides the intercept parameter (events per month) with standard error, and GLM slope parameter (events per year) with standard error and p-value. These statistics are given for events occurring to US firms (US), to non-US firms, and to all firms together (All). Statistics were taken on the window of January 1^{st}, 2007 until April 15^{th}, 2015. For each of the aforementioned categories, the statistics are given for all sizes (including 2683 events with unknown size within the US, and 526 events with unknown size outside of the US), and for events with breach in excess of fifty thousand ids. ($> 5 \times 10^4$).

7.2.1 Data Breach Frequency

The rate (frequency) of breach events is studied, with relevant statistics presented in Tab. 7.1. According to a linear regression of monthly counts over time (Poisson Generalised Linear Model with identity link [195],) the rate of large events has been stable within the US, and growing significantly outside of the US – driving almost significant ($p = 0.13$) growth when all countries are taken together. However, this growth is 0.18 events per year, which is only a fifth of a percent of the total annual rate, thus being practically insignificant. This apparent stability runs counter to the view that cyber risks are worsening. Next, we consider the dynamics in the size of large breaches, which provides a less reassuring message.

7.2.2 Data Breach Severity

The dynamics of the df of large breach sizes over time are studied. Given the growing amount of data being stored online, and the evolution of cyber crime methods, the severity of breaches is expected

to increase. As an initial diagnostic, the observed cumulative sum of large breaches over time is shown, in Fig. 7.1 (panel IV), to curve up – indicating growth in the mean breach size. We thus consider three possibilities: (i) there is a maximum possible breach size, (ii) this maximum is growing, and (iii) large breaches are becoming larger over time.

Detecting The Maximum Breach Size

We use Extreme Value Theory (EVT) [79] to determine if a maximum breach size exists, and if so, how it has evolved. For this we assume that the breach counts are large enough that they can be treated as realizations of a random variable with a continous rather than discrete distribution. EVT provides a standard framework for such statistical inference "beyond the largest observation". We exploit a quite general EVT theorem that roughly states: for large values of random variable X being in excess of a sufficiently large truncation u, the *Generalized Pareto Distribution* (GPD) approximately models the tail of the df of X,

$$Pr\{X - u \leq x | t,\ X - u > 0\} \approx 1 - (1 - \xi x/\beta(t))^{-1/\xi}\ ,\ \text{with}\ \ \beta(t) > 0\ \ ,\ \text{for}\ \ \xi \neq 0\ , \tag{7.1}$$

with $0 \leq x$ without upper bound if the *Extreme Value Index* ξ is positive and $0 \leq x \leq -\beta(t)/\xi$ for $\xi < 0$. In this latter case, a finite maximum exists for X,

$$\nu(t) = u - \beta(t)/\xi < \quad \infty\ , \tag{7.2}$$

which may vary with time when the scale parameter β is time dependent.

To characterize the maximum breach of the (natural) log of breach sizes, $\nu(t)$, we consider the following statistical hypotheses and their corresponding parameterisations of (7.1),

$$M_0: \text{ no maximum size is detected } (\xi > 0)\ ; \tag{7.3}$$

$$M_1: \text{ there is a constant log-maximum } (\xi < 0,\ \beta(t) = \beta_0)\ ; \tag{7.4}$$

$$M_2: \text{ the log-max grows linearly in time } (\xi < 0,\ \beta(t) = \beta_0 + \beta_1 t,\ \beta_1 > 0)\ ; \tag{7.5}$$

$$M_3: \text{ the log-max grows sub-linearly in time } (\xi < 0,\ \beta(t) = \beta_0 + \beta_1 \ln(t),\ \beta_1 > 0)\ . \tag{7.6}$$

For the GPD approximation (7.1) to be good, one wants u to be as large as possible, but at the same time one wishes to have a large sample. Thus, we take estimates at the lowest value of u at which parameter estimates stabilize. The GPD (7.1) with its parameterisations (eqs. 7.4-7.6) were estimated on the natural log of the breach sizes (*Peaks Over Threshold (POT) Estimation* [54]), for lower thresholds ranging from $u = 14.4$ (having 102 points above) to $u = 16.8$ (having 20 points above). Considering that the estimated maximum is approximately stable for $u > 15.5$ (Fig. 7.1, panel (II)), estimates are taken at $u = 15.5$ (having 50 points above), and recorded in Tab. 7.2.

The main insights about the behaviour of the maximum size gained from this estimation are visualized in Fig. 7.1 (panels (I) and (II)). For a range of lower thresholds u, the estimated maximum breach size are "hugging" the data, implying the existence of a highly significant upper truncation which the data has already reached. Moreover, we find a highly significant upward growth of this maximum size. In further detail, for M_1, the estimated shape parameter ξ is already significantly negative ($\widehat{\xi} = -0.36\ (0.1),\ p \approx 0.001$) with small $u = 14.4$, and achieves values below -1 for $u > 16$, indicating a highly significant maximum. Both M_2 and M_3 exhibit significant growth in the maximum over time ($p < 0.001$ in Tab. 7.2), and have superior likelihood to M_1 by the likelihood ratio test (LRT), having $p = 0.08$ and $p = 0.05$, respectively. Finally, M_3, has superior log-likelihood to M_2 for all $u > 15.5$ with the same number of parameters. Thus, the best model for the growth of the maximum (obtained by exponentiation of the log-maximum) is the sub-linear power,

$$\exp(\nu(t)) \propto t^{-\beta_1/\xi} \approx t^{0.83}. \tag{7.7}$$

Indeed it may seem obvious that here – as in any natural (finite) system – there is a maximum size. Further, given the flow of users and data online, and the growth of giant IT companies, it is sensible that this maximum possible breach size is increasing. However, to quantify this is of high importance to policymakers and (re-)insurance firms who care about the aggregate risk, which is dominated by the largest observations of such heavy tailed df. It also makes a major theoretical distinction as for this model with a finite maximum all moments are finite, whereas with an infinite maximum all moments are infinite.

Table 7.2: EVT peak-over-threshold (POT) estimates of the three models (eqs. 7.4-7.6) are presented for lower threshold u, with loglikelihood (ll), and estimated shape ξ (with standard error), scale intercept β_0 (with standard error) and scale slope β_1 (with standard error and p-value).

Model	u	ll	ξ	β_0	β_1
M_1	15.5	-60	-0.61 (0.18)	2.24 (0.47)	$= 0$
M_2	15.5	-57	-0.65 (0.15)	1.35 (0.39)	0.18 (0.06), 0.001
M_3	15.5	-56	-0.78 (0.15)	1.60 (0.26)	0.65 (0.15), 0.7×10^{-6}

The Distribution of Large Breach Sizes

The df of large breach sizes is estimated to quantify the severity of data breach risks and their dynamics. We model the large breach sizes by a doubly truncated Pareto (DTP) df typical for modeling extreme risks,

$$F_{DTP}(x|t) = \frac{1-(x/\widetilde{u})^{-\alpha(t)}}{1-(\widetilde{\nu}/\widetilde{u})^{-\alpha(t)}} \ , \quad 0 < \widetilde{u} < x \leq \widetilde{\nu} \ , \ \alpha(t) > 0 \ , \tag{7.8}$$

having *shape parameter* $\alpha(t)$ potentially varying in time. Rather than working directly with (7.8), for convenience we work with the natural logarithm of the data, $Y = \ln(X)$, which follows a doubly truncated Exponential (DTE) df,

$$F_{DTE}(y|t) = \frac{1-\exp(-\alpha(t)(y-u))}{1-\exp(-\alpha(t)(\nu-u))} \ , \quad 0 < u < y \leq \nu \ , \ \alpha(t) > 0 \ , \tag{7.9}$$

with $u = \ln(\widetilde{u})$ and $\nu = \ln(\widetilde{\nu})$.

We have already determined that a significant and growing maximum breach size exists. However, it is not yet known what other trends have been present within breach severity. We thus consider statistical hypotheses about trends in the df of large breaches, and their corresponding parameterisations in (7.9):

D_0 : the df has a fixed max ($\nu(t) = \nu_0$) and is stationary ($\alpha(t) = \alpha_0$) ; (7.10)

D_1 : the df has a fixed max ($\nu(t) = \nu_0$) and becomes more heavy tailed ($\alpha(t) = \alpha_0 + \alpha_1 t \ , \ \alpha_1 < 0$) ; (7.11)

D_2 : the maximum log-breach grows sub-linearly ($\nu(t) = \nu_0 + \nu_1 \ln(t) \ , \ \nu_1 > 0$),
and the df becomes more heavy tailed ($\alpha(t) = \alpha_0 + \alpha_1 t \ , \ \alpha_1 < 0$) . (7.12)

The hypothesis D_2 (7.12) contains (7.6) where $\nu_0 = u - \beta_0/\xi$ and $\nu_1 = -\beta_1/\xi$. The hypotheses

D_0 (7.10) and D_1 (7.11) overlap with testing of the previous hypothesis tests (7.3-7.6), providing an opportunity for confirmation of results about the maximum with the specific DTE model (7.9).

The DTE (7.9) with specifications (7.10-7.12) were estimated by Maximum Likelihood (ML). For D_2, the maximum was given by $\nu_0 \approx 13.63$ and $\nu_1 \approx 0.84$, computed from the EVT estimates (M_3 in Table 7.3).

The estimation (Table 7.3) finds that large breaches have grown larger (the tail has become heavier), and confirms that a maximum breach size exists, and is increasing. More specifically, it is confirmed that a maximal breach size is clearly present by comparing the likelihood of D_0 with and without finite maximum ($p \approx 10^{-7}$ by the LRT). It is also found that large breaches have become heavier tailed, with the shape parameter of D_1 significantly decreasing ($p = 0.04$) at a rate of -0.027 per year. This extension of the model significantly increases the quality of fit (D_1 has superior likelihood to D_0 with $p_{LRT} = 0.006$). Finally, in confirmation with the EVT results, the model (7.12) is found to be best, where the growing maximum further improves the result of D_1 (D_2 has superior likelihood to D_1 with $p_{LRT} = 0.007$).

These results are all quite striking. For instance, all estimates, having shape parameter $\alpha < 1$, are so extremely heavy tailed that, without a finite maximum, the mean of this model would be infinite. For D_0, having shape parameter $\alpha = 0.47$ and maximal breach 1.52×10^8, the expected large breach size is 3.1×10^6, with even larger standard deviation, 1.3×10^7 (7.14). For D_2 in 2015 ($t = 8$), having $\alpha(8) = 0.364$ and maximum $\exp(\nu(t)) = 2.24 \times 10^8$, the expected large breach size becomes twice as large as for D_0. Further, under D_2, given a breach size is in excess of fifty thousand, there is a 10 percent chance that it exceeds ten million ids.

In Fig. 7.1 (panel III), diagnostics are given to demonstrate the validity of D_2. For this, the data are transformed to be stationary and then standard diagnostics are performed. In detail, if D_2 is true, then the log breach random variable $Y = \ln(X)$ is from model (7.9) with non-stationary parameters (7.12) with estimated values in Tab. 7.3. Thus, the transformed data,

$$\widetilde{Y}_i = Y \times \frac{\alpha(t)}{\alpha_0} \mid t \sim \mathrm{DTE}\left(\alpha_0\ ,\ \nu^*(t)\right)\ (7.9), \qquad \nu^*(t) = \frac{\alpha(t)}{\alpha_0} \times \nu(t) \approx \nu_0\ , \tag{7.13}$$

will be approximately stationary with df equal to that of the log breaches Y at time 0 (January 1^{st}, 2007). Estimating the model (7.9), with the D_0 specification, on the transformed data (7.13), yields an estimate of $\widehat{\alpha} = 0.58$, which agrees with the estimate D_2 at $t = 0$ ($\alpha(0) = \alpha_0 = 0.57$), indicating that

the transformation is valid. Further, in Fig. 7.1 (panel III), the empirical complementary CDF of the transformed data are found to be well described by this estimated stationary model, as evidenced by small residuals between the empirical and estimated complementary CDFs – the Kolmogorov-Smirnov [268] has p-value of 0.78 – as well as the consistency of the shape parameter α for lower truncations u ranging from 5×10^4 to 5×10^7.

The above results are strengthened and confirmed by obtaining similar results with a more flexible method. We used quantile regression [153] which, rather than specifying a parametric model, independently fits linear regressions to each quantile. In support of the dynamic specifications D_1 and D_2, Fig. 7.1 (panel I) shows good agreement between the linear quantile regressions and the growing quantiles of D_2. Further, the slope, standard error, and p-value of the linear regressions of the 0.5 and 0.9 quantiles are 0.083 (0.038), 0.03 and 0.145 (0.07), 0.04, respectively. Thus the quantile regressions are significantly increasing, providing strong additional evidence that large breaches are getting larger.

Table 7.3: Parameter estimates, standard errors (Monte Carlo with 1000 repetitions), and p-values for the significance of slope parameters (Monte Carlo with 1000 repetitions), of model (7.9) with parameter specifications eqs. 7.10-7.12. The D_0* model has no finite maximum.

Parameter	ll	α_0	α_1	ν_0	ν_1
D_0	-1020.7	0.47 (0.017)	$= 0$	18.839 (0.2)	$= 0$
D_0*	-1032.9	0.51 (0.02)	$= 0$	$= \infty$	$= 0$
D_1	-1017.0	0.58 (0.05)	-0.027 (0.01), 0.002	18.839 (0.2)	$= 0$
D_2	-1013.3	0.57 (0.05)	-0.025 (0.01), 0.014	$= 13.63$	$= 0.84$

7.2.3 Cumulative Risk & Future Projections

Due to the large-scale sharing of breached data, e.g., by sophisticated underground markets [99, 207], breached personal information is concentrated, enabling efficient subsequent identity fraud [113]. Thus, privacy *erodes* with the growth of the cumulative number of breached ids. For this reason, to understand the long term risk of data breaches to privacy, it is crucial to study the cumulative amount of breached information. The cumulative sum measure brings together both frequency and severity in a convenient way for the study of past and future evolution of risk.

In Fig. 7.1 (panel IV), the observed cumulative sum, $C_n = \sum_{i=1}^{n} x_i$, of the $n = 619$ large breaches occurring from January 1^{st}, 2007 until April 15^{th}, 2015, is plotted. If both the statistics of frequency

and severity were stable over time, the cumulative sum would grow linearly in n. That the observed cumulative sum curves upward indicates a growing mean, as featured in the D_1 and D_2 specifications of (7.9). To compare the observed data with the models, the expected cumulative sum $E[C_n] = nE[X]$, and its standard deviation, are plotted for the estimated models. This simply requires computing the first two moments of the DTP (7.8):

$$\mathrm{E}[X] = \frac{\alpha}{\alpha - 1}\left[\frac{u^{1-\alpha} - \nu^{1-\alpha}}{u^{-\alpha} - \nu^{-\alpha}}\right] \ , \quad \mathrm{E}[X^2] = \frac{\alpha}{\alpha - 2}\left[\frac{u^{2-\alpha} - \nu^{2-\alpha}}{u^{-\alpha} - \nu^{-\alpha}}\right] \ . \tag{7.14}$$

However, there is an important subtlety in comparing the observed and expected curves. The relevance of the expected sum to the observed sum relies upon the Central Limit Theorem (CLT). For infinitely large ν, the mean is infinite and thus the CLT never converges. In this case, the cumulative sum would grow super-linearly, $C_n \sim u\, n^{1/\alpha}$, rather than linearly [244]. For instance, for $\alpha = 0.5$, this curve would have exceeded the upper boundary of the Fig. 7.1 (panel IV) within the first $n = 200$ observations. A rule of thumb for when the CLT holds, and the cumulative sum grows linearly in n rather than super-linearly, is for samples larger than $n^* \approx (\frac{\nu}{u})^\alpha$ [244]. Here where $u = 5 \times 10^4,\ \nu = 1.6 \times 10^8$ and $\alpha \approx 0.5$, the crossover point is $n^* \approx 50$. This means that the observed upward bend in the cumulative sum occurring after n^* is not due to the superlinear growth resulting from the heavy-tailed CDF but rather results from the non-stationarity of the process. Comparing the observed and expected curves: D_0, which grows linearly, fails to capture the trend; D_1, which is convex due to an increasingly heavy tail, partially captures the trend; and D_2 is again the best model, capturing the trend well due to both an increasing maximum and increasingly heavy tail.

For future projections, we make use of all fitted models. The previous analysis was conditional upon knowing the number of breaches. To more properly quantify the uncertainty of projections, the annual number of breaches N_t is treated as random. The annual sum,

$$Y_t = X_1 + \cdots + X_{N_t} \ , \tag{7.15}$$

(with all X_i and N_t independent) is called a *Compound Process* [182, 296]. The mean and variance are given by,

$$\mathrm{E}\left[Y_t\right] = \mathrm{E}\left[X_t\right]\mathrm{E}\left[N_t\right] \quad , \quad \mathrm{Var}(Y_t) = \mathrm{E}\left[N_t\right]\mathrm{Var}(X_t) + \mathrm{E}\left[X_t\right]^2\mathrm{Var}(N_t) \quad , \tag{7.16}$$

which are computed for the coming five years of 2015-2019 in Tab. 7.4, where the four months of data

after January 1^{st}, 2015 were excluded for convenience. One observes that the cumulative breach C_t, currently at a level of around 1.816×10^9, is expected to more than double in the next five years under the non-stationary D_1 and D_2 model specifications.

Model	Quantity	2014	2015	2016	2017	2018	2019
D_0	$Y_t \times 10^{-8}$	-	2.37 (1.14)	2.37 (1.14)	2.37 (1.14)	2.37 (1.14)	2.37 (1.14)
D_0	$C_t \times 10^{-8}$	$= 18.16$	20.5 (1.14)	22.9 (1.61)	25.3 (1.94)	27.6 (2.28)	30.0 (2.55)
D_1	$Y_t \times 10^{-8}$	-	3.73 (1.52)	4.18 (1.63)	4.67 (1.74)	5.21 (1.86)	5.81 (1.99)
D_1	$C_t \times 10^{-8}$	$= 18.16$	21.9 (1.52)	26.1 (2.22)	30.7 (2.82)	36.0 (3.38)	41.8 (3.92)
D_2	$Y_t \times 10^{-8}$	-	4.96 (2.12)	5.97 (2.48)	7.16 (2.87)	8.56 (3.31)	10.2 (3.79)
D_2	$C_t \times 10^{-8}$	$= 18.16$	23.1 (2.12)	29.1 (3.27)	36.3 (4.35)	44.8 (5.47)	55.0 (6.65)

Table 7.4: Future expected annual breaches (and standard deviation) (7.16) of the annual sum Y_t (7.15) and cumulative breach C_t since 2007 are presented with annual rate estimates from Table 7.1, and DTP df (7.8) with parameterisations given in eqs. 7.10-7.12, and parameter values in Table 7.3. All values in the table are divided by 10^8. These estimates assume a constant rate of 75.6 large events per year with annual variance of 229 (computed across 2007 until 2015). The US rate estimate is 50.4, and thus US estimates may be obtained by scaling the numbers by $49.5/75.6 \approx 2/3$ (the US total is 11.89×10^8, which is also about 2/3 of the total of 19.62×10^8).

7.2.4 Data Breach Risk & Organisation Size

Larger organisations tend to have more attractive assets to motivate a cyber attack, present a larger *attack surface* making penetration easier, and once penetrated, contain more data to be exfiltrated [233]. Thus, larger organisations will be both more frequently and severely victimised than smaller ones. To quantify the relationship between risk and organisation size, we use market capitalisation (MCAP) as a proxy for size, and compare organisations of different size with past observed breaches. Specifically, we consider 4,950 firms publicly traded on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and the Nasdaq [187]. For victimised firms, MCAP was taken the day before the data breach incident, to avoid possible subsequent devaluations due to the breach (e.g, see [43]), and to avoid price changes that often occur in the highly dynamic stock market. For non-victimised firms, the MCAP was taken at June 1^{st}, 2014. All MCAP values were then inflation adjusted to 2014 US Dollars. Within the dataset studied here, there were 735 events across 400 of these companies, between 2006 and 2015.

To model the cumulative distribution function (CDF) of the size of all firms ($n = 4,950$), and the size of firms given that they have been victimised ($n = 735$), we consider a Lognormal df. This

model choice has substantial justification, being a good model for a variety of size measures ranging from profits, to MCAP, to the number of employees [267, 173, 107] and because it asymptotically encompasses the Zipf law family [175] often advocated based on the interplay between proportional growth and firm birth and death events [251, 240, 239, 216]. The df is truncated from above and below by the observed maximum and minimum of 10^6 and 6.6×10^{11}, respectively. The estimated dfs are plotted with the empirical dfs in Fig. 7.2 (panel (I)), demonstrating reasonable goodness of fit. The CDF of victims, being shifted about 1 decade to the right of its unconditional counterpart, gives much higher probability to larger firms being victimised. This is further studied in the next subsection, and the relationship between firm size and breach size is presented afterwards.

Breach Frequency & Firm Size

The effect of firm size on the frequency of breach events is now studied. Despite the fact that only 10 percent of the publicly traded firms have been victimised, nine of the ten largest have been, and often multiple times (e.g., both Apple Inc. and IBM 6 times, Facebook 8 times, Bank of America 8 times, HSBC 17 times, General Electric 2 times, and Wal-Mart 6 times). Thus larger firms, despite being rare, are frequently victimised. The relative frequency at which firms with a given log(MCAP) are victimised is quantified by the victimisation pdf,

$$\Pr\{\text{Firm log(MCAP) size}= x|\text{Firm is victimised}\} \propto \frac{f_{\text{Victim}}(x)}{f_{\text{Population}}(x)} \quad , \tag{7.17}$$

which is proportional to the ratio of the victim and population log size densities. Using the Lognormal estimates from the previous section as well as histogram density estimates, the victimisation pdf (7.17) is plotted in panel (II) of Fig. 7.2. Both the histogram and parametric estimate suggest approximate linear growth, with slope 0.6, for sizes between 10^8 and 10^{11}. Thus the probability of victimisation grows with size as $\sim s^{0.6}$.

The way that frequency (and breach size, to be seen below) scale with firm size can be interpreted as a quantification of an effective *risk surface*. A 3D object with volume c has surface area proportional to $c^{2/3}$. In d dimensions, the exponent is $(d-1)/d$. Thus the exponent 0.6 would correspond to a (fractal) dimension of $d_f = 2.5$, implying that the risk surface scales with firm size in a way between that of the circle delineating a disk and the sphere bounding a ball.

Another line of interpretation acknowledges that firms are composed of sub-units, and that

breaches occur at this level. An example of this is Sony Corporation, which suffered a data breach on its PlayStation Network (Sony Computer Entertainment Division) in 2011, and a separate massive attack on Sony Picture Entertainment in late 2014. In [15], it was found that the number of subunits in a firm scales with firm size as $\sim s^{1/3}$ and that the size of the largest sub-unit within a firm scales like $\sim s^{2/3}$. A pleasant, but simplistic, connection between this scaling and our result would be that the probability of attacks is proportional to the size of the largest unit, which is arguably the most visible and vulnerable. In reality, the number of subunits may also play a role.

Breach Size & Firm Size

We now quantify the way in which breach size scales with firm size. Companies with larger market capitalisation, tending to have more customers, retain more personal data that can be leaked. Further, personal data are increasingly considered as the *ore* that companies *mine* to extract consumer profiles and enhance online commerce [233]. Thus, the value (and size) of personal data assets are increasingly likely to be reflected in market capitalisation [44].

In Fig. 7.2 (panel III), the 298 breaches with size in excess of 10^3, that occurred at publicly traded US firms, are plotted versus organisation size (MCAP). Further, linear quantile regressions of this data, with automatically detected change-points (additive quantile regression [153]), are plotted. One sees that the largest breaches occur at larger organisations. That is, the quantile regressions for quantiles 0.5 and above significantly increase ($p < 0.05$) with firm size, until this size effect saturates for firm sizes above 10^{10}. In particular, for the 0.9 quantile, breach size increases with slope 0.66, indicating that the largest breach sizes scale with firm size as$\sim s^{0.66}$.

It is also apparent that, despite the fact that large firms suffer the largest breaches, they also – more frequently – suffer small ones in which only a fraction of the total information assets are extracted. Assuming that hackers always aim to maximise the volume of ex-filtrated information, these small breaches could be considered as only partially successful. More likely, again recognising that organisations are comprised of functional sub-units – potentially having separate IT systems – and that attacks occur at the sub-units, then the breach sizes will be related to the subunit sizes. Again recalling the result of [15] that the size of the largest sub-unit scales with firm size like $\sim s^{2/3}$, we see an incredible coincidence that the 0.9 quantile of breaches and breach frequency also scale with firm size in this way. This strongly links the risk of breaches to the size of the largest sub-unit.

This characterization of the risk surface in terms of firm size – specifically with densities that dictate both how the number of breaches are distributed across firms of different size, and how large breaches tend to be for victimized firms of different size – allow for the extrapolation of breach statistics onto unobserved populations. For instance, given a population of organizations and their sizes, one can infer the quantity and distribution of breaches that they have suffered. The validity of this most basic extrapolation will involve a ceteris parabis assumption. For instance, one will have to assume that such an organization is similar in attractiveness and permeability to cyber criminals to a publicly traded firm of similar size.

7.2.5 Sector & Data Breach Risk

The previous sections focused on identifying universal relationships in data breach risk with a single risk factor, organisation size. Clearly, there are other attributes of organisations that are relevant to data breach risk. We thus consider industrial sector as a risk factor, which may serve as a proxy to identifying relatively homogeneous subpopulations in regard to their frequency of interaction with consumers, and the total volume of personal data that they guard. There are twelve sectors, as defined in [187]. The sector is likely confounded with firm size, and thus firm size effects. However, due to the limited data, a statistical analysis considering these effects jointly is infeasible. Rather, a simple analysis is done, nuancing the former results.

Fig. 7.3 plots the median and 10 year frequency of large breaches for the 12 sectors. We propose to rationalise these observations by the hypothesis that the frequency of breaches is related to the frequency of customer interaction, and that the severity should be related to the volume of personal data guarded by an organisation. Consumer Service, and Finance have small sub-units (i.e., retail shops and bank branches) at which they interact with local customers, and suffer from small but frequent breaches. On the other side, Basic Industries, having large centralized operations, and infrequent customer interaction, implies large yet infrequent breaches. Consumer Durables have a lower breach frequency than Non-Durables, which, by definition, involve more frequent customer contact. Further, the companies with the largest loss in the three categories with highest median loss are Sony (Non-Durables), eBay (Misc.), and UPS (Transportation). These companies all clearly guard large amounts of personal data. Capital Goods (e.g., heavy equipment producers) suffer relatively small breaches, possibly due to a smaller number of customers than retailers of non-durables. These comparisons tend

to support the hypotheses posed above.

7.3 Discussion

Due to the combination of large size and the high potential for immediate financial damage, breaches of personal identities (ids) are among the most disruptive and costly cybersecurity events both for consumers [113] and organisations [235]. Thus, large breaches translate into immediate financial consequences for organisations, often reflected in drops in stock price [43, 100, 8, 101], and reputational damage. Based on this study, the annual total of breached ids is expected to range between half a billion to a billion over the next 5 years. Considering the average cost of more than 200 USD per breached item [202], this translates into hundreds of billions of USD losses per year. Severely worsening the problem is that, not only do data breaches have short term consequences for individuals, but due to sophisticated underground markets for breached data [99, 207], breached personal information is concentrated, enabling efficient subsequent identity fraud [113]. As an illustration, several research studies have established how personal identities may be reconstructed, and individuals re-identified from a few spatio-temporal locations of credit card uses [68] or even from public data [9]. Thus, privacy gets increasingly *eroded* as the cumulative number of breached ids grows.

Cyber risks, within the general context of the evolving and expanding Internet, are highly dynamic. Being governed by the struggle of IT security technology to keep up with the constant innovation and adaptive nature of cyber crime, ranging from social engineering attacks to the egregious sale of "zero day" security vulnerabilities. The struggle is compounded by an ever-growing amount of personal data stored online, and a growing attack surface due to the adoption of mobile computing paradigms. Due to the clear potential for damage, it is crucial that the risk of data breaches be well understood. Despite the dynamic context of data breaches, and a relatively short history, we have specified a statistical model for risk that successfully unearths clear statistical regularities, allowing for novel understanding of this risk.

The frequency of large breaches (having in excess of fifty thousand ids) was found to be stable since 2007, with a rate of 76 events per year. Despite the relatively low rate, since 2007 nearly 2 billion pieces of personal information have been breached, hinting at the extreme size of large breaches. For breach sizes, we considered three possibilities: (i) there is a maximum possible breach size, (ii) this maximum has grown, and (iii) large breaches became larger over time. These statements were formally

tested on the data (from January 1^{st}, 2007 until April 15^{th}, 2015). By Extreme Value Theory (EVT), it was found that a highly significant maximum breach size exists, and is growing with time like $t^{0.84}$, where the current maximum breach size is about 200 million, and is expected to grow fifty percent over the next five years. This feature is highly relevant for policymakers and re-insurance companies who are concerned with evaluating and managing the total risk.

On a critical note, this upper limit may not be entirely valid. That is, EVT cannot consider the presence of an unobserved regime beyond the *usual* df. Such singular extreme statistics, also called *Dragon-Kings*, have been observed in, e.g., city sizes, financial crashes, and nuclear accidents [259, 291]. For example, Facebook gathers personal data from 1.4 billion users (as of January 2015, and not counting Whatsapp (700 million users) acquired by Facebook in 2014), and the National Security Agency (NSA) most probably gathers personal information about several billion people worldwide. Though these organisations certainly operate with incomparably more resources, and at a much high level of information security, the chance of massive attacks on personal data gathered and stored by Internet giants and governments cannot be excluded in the future – especially considering the size effect by which large organisations are more frequently attacked, discussed below.

The df of breach sizes was found to be well modeled by a Pareto (Power law) df with a linearly shrinking shape parameter, and maximum breach size given by the EVT estimate. The estimate of this model, having shape parameter 0.57 (0.05) in 2007, is expected to have become much heavier tailed with shape parameter 0.37 in 2015. Under this current model, given that a breach is in excess of fifty thousand ids, there is a ten percent chance that the breach exceeds ten million ids.

Next, the connection between organisation size, measured by market capitalization, and the risk of data breaches was unearthed. It was found that the frequency of breaches scales with organisation size like $\sim s^{0.6}$, indicating that larger organisations are victimised at much higher frequencies than their smaller counterparts. Further, the largest breach sizes (quantified by the 0.9 quantile) were found to scale with organisation size $\sim s^{0.66}$ for firms with market capitalisation between one million and ten billion US$. Above ten billion US$, the scaling reached a plateau. Thus we identified that the effective *risk surface* scales with organisation size with at exponent around 0.6. This can be thought of as a fractal scaling relationship. Alternatively, recognising that organisations are comprised of functional sub-units – potentially having separate IT systems – and that attacks occur at the sub-units, then the breach sizes will be related to the subunit sizes. In [15], it was found that the size of the largest

sub-unit scales with firm size like $\sim s^{2/3}$. That this scaling relationship is similar to that of the risk surface, is highly suggestive that the size of the largest sub-unit – potentially housing the main IT system and data – defines the risk level. This is a precious first step towards establishing a relationship between data breaches and the underlying structure of organisations in which they tend to occur.

As the damage of data breaches is a cumulative erosion process, we also studied the cumulative sum of data breaches. Like many negative externalities in the economy [275], the phenomenon of privacy erosion is hard to measure. For instance, the cumulative sum likely overestimates the erosion as, to some extent, the ex-filtrated data degrades over time. In particular, users can change their passwords and cancel their credit cards. However, other aspects of the victims identity – such as name, address, and social security number are more persistent. Thus, in reality, the true erosion is somewhere in between an instantaneous and cumulative process. However, this approach provides a transparent quantification upon which discussions of past and future risk may be had. To date, the total amount of breached personal data, well approximated by the sum of large breaches, is around 2 billion personal id items. Although some individuals may have been subjected to multiple breaches, the amount of aggregate loss is considerable, e.g., compared to the number of Internet users (3.5 billion) or even to the world population (7 billion). Projections based on the best models suggest that the growth of the cumulative sum is accelerating, and is expected to double (surpassing the number of Internet users) in the next 5 years. Finally, it is important to note that about 30 percent of events have unknown breach size, and thus the statistics above are underestimates.

Our results provide detailed insights on the aggregate statistics of data breaches, their impact on organisations, as well as on the long-term effects for consumers and society. In fact, the provided estimates of the risk level are conservative given the fact that we are only considering the subset of data breaches that have been discovered, reported, and submitted to the online databases. Our findings suggest that people should not only worry about privacy erosion the way we experience it nowadays, by assuming that our personal data will remain confined and exploited only by "trustworthy" private organisations and governmental agencies. On the contrary, people should expect that the personal data they have uploaded on online platforms, such as social networks, may be either suddenly or progressively disseminated, first in underground markets, and then publicly.

Finally, this work supports and encourages further emphasis on addressing cyber risks at both organisational, and governmental levels. This risk should not be thought of in terms of representative

or typical data breaches, which is a completely inadequate concept given the heavy-tailed nature of the pdfs. Rather one must consider the full distribution. Moreover, this is a system where the total risk is dominated by the few extremes, which are inherently stochastic. Thus, policies should be adapted to this regime where standard actuarial methods, insurance by the mean, do not apply. In terms of defense, this also calls for measures aimed at stopping the big events, the cascades, perhaps by decentralizing IT systems in big organisations.

7.4 Future work

The findings of this study are useful and important for many stakeholders in cyber risk. However, this study leaves many areas unexplored, and also identifies important areas for future study:

- *Extending scope by extrapolation:* The data is largely limited to English speaking countries, and is most reliable for the United States, where data is provided by the government based on the Freedom of Information Act. Thus, to arrive at a more complete picture of the risk, new data sources for other countries may be considered. However, it is unlikely that such data will be made available as few other governments allow free access to such data, and have less strict laws about reporting breaches. So, instead, we propose to perform an extrapolation based on auxiliary information. For instance, having discovered the relationship between firm size and breach risk, we can consider a population of firms whose breaches have been unobserved, but whose firm sizes are known, and infer the quantity and distribution of breaches that they have suffered. This can be made more sophisticated by considering other factors in addition to firm size.

- *Further exploring organization size and risk:* Market capitalization was used to define firm size. Thus, only publicly traded firms were considered. Most of the victims are private firms. Thus number of employees, or another measure, should be used to allow for the consideration of all firms. Furthermore, as was suggested by the study, risk scales with size as $s^{0.6}$, which is also how the size of the largest organizational sub-unit scales with size [15]. Thus, the apparent connection between risk and sub-unit size should be tested by comparing sub-unit size data (as in [15]) with breach data.

- *More specific risk with more risk factors:* The study considered firm size as a risk factor, and suggested some link between sector and the degree of customer interaction that a firm has. The

rigorous quantitative study of risk factors aside from firm size should be done, such as sector, the type of data that a firm tends to have, and its degree of interaction with customers. This will allow for more specific discussion of risk, and the identification of what makes certain firms more attractive and vulnerable to attacks.

- *Linking the evolution of risk to the evolution of the Internet:* The study made no quantitative connection to how cyber risk scales with the growth and evolution of the Internet. The number of people with Internet connectivity, the number of users of social media, and the amount of personal data going online should be considered. Thus, to better understand the evolving risk of data breaches, it is clear that the link between the past dynamics of breaches, and the growth of the Internet be quantified. Aside from being interesting in itself, this will allow for more informed forecasts to be made.

- *Keeping up to date:* The Internet, technology, cyber crime, and the amount of personal data online are evolving. Thus, the data and the risk models should be dynamic and kept up to date. That is, it is important that the study, and extensions of it, be performed regularly to provide the most accurate assessment of the present and future risk.

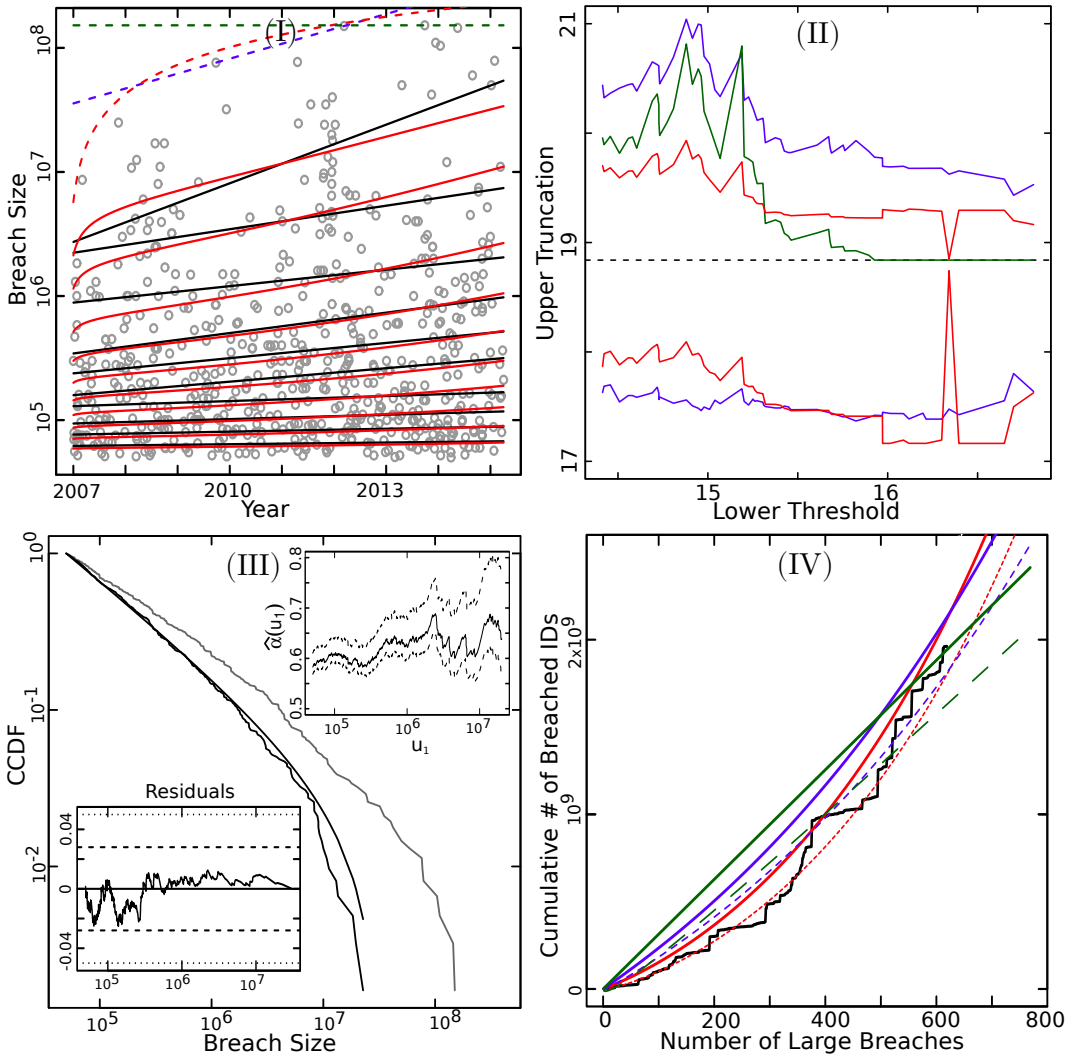

Figure 7.1: Panel (I) plots large events (above 5×10^4) over time from January 1^{st}, 2007 until April 15^{th}, 2015, as well as the $p = (0.1, 0.2, \ldots 0.9)$ and 0.95 quantile levels of a linear quantile regression (black), and D_2 (eq. 7.8; red), estimated on the data. The upper endpoints of M_1 (green), M_2 (blue) and M_3 (red) are given by the dashed lines. Panel (II) plots EVT estimates of the maximum of the natural log of the data. M_1 is given in green. M_2 (blue) and M_3 (red), having growing maximum, have the maximum plotted at both January 1^{st}, 2007 (lower) and at April 15^{th}, 2015 (upper). Panel (III), in the main frame, plots the empirical complementary cumulative distribution function (CCDF; rough black line) of the transformed breach sizes (for the second alternative model), the DTP CCDF (smooth black line) $\widehat{\alpha} = 0.56$) estimated on the transformed breach sizes, and the empirical CCDF for the untransformed data (grey line). The inner left frame plots the "residual" distances between the empirical and estimated CDFs (the two black lines), with $(0.1, 0.25, 0.5, 0.75, 0.9)$ quantiles plotted for the Kolmogorov Smirnov df. The inner right frame is a sequence of DTP shape parameter estimates, and standard errors, on the transformed data, for increasing lower truncations, u. Panel (IV) plots the observed cumulative sum of the 619 large breaches occurring from January 1^{st}, 2007 until April 15^{th}, 2015 (rough black). The horizontal axis extends to 800, which is the number of events to be expected about 2.5 years into the future. The green, blue, and red lines provide the expected cumulative sum under the estimated null, first, and second alternative models, respectively. The lower standard error for each of the three curves is plotted in the same color, with the null, first, and second alternative models having long dashed, medium dashed, and densely dashed lines respectively. The upper standard error will be the same distance from the expected cumulative sum, but on the positive side.

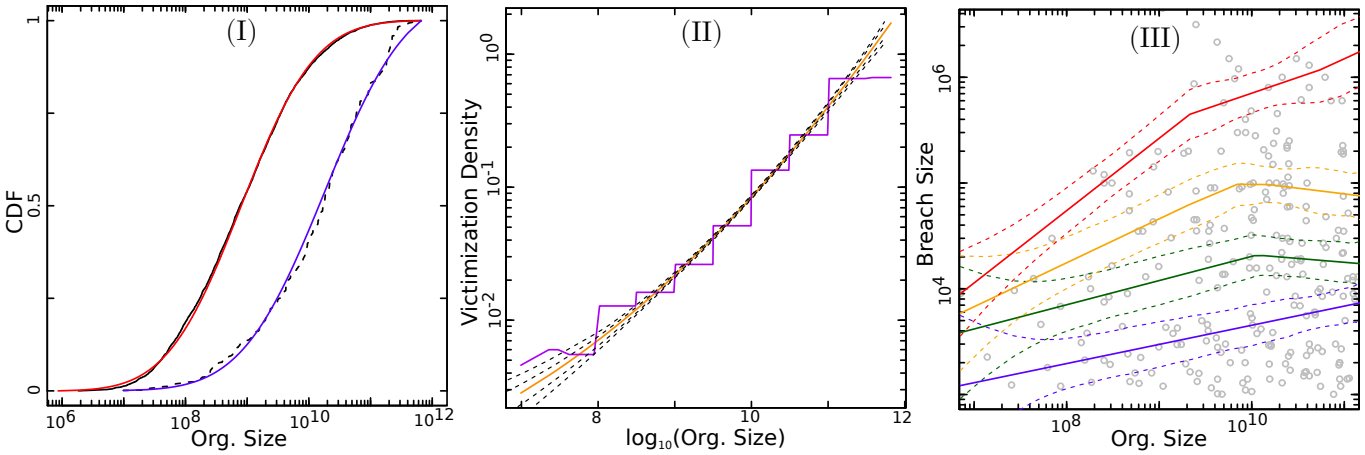

Figure 7.2: The effect of firm size (market capitalisation) on the risk of breaches: Panel (I) provides the empirical (black) and estimated Lognormal CDFs for organisation sizes (red), and victimised organisation sizes (blue). The estimated parameter values and standard errors are ($\mu = 20.3$ (0.06), $\sigma = 2.1$ (0.06)) and ($\mu = 23.6$ (0.14), $\sigma = 2.5$ (0.1)), respectively. Panel (II) plots the estimated pdf defining the probability that firms of different log-size are victimised. The purple kinked line is computed by taking histogram estimates of the population and victim log-size densities and plugging them into (eq. 7.17). The orange line does the same but with the Lognormal estimates. Monte Carlo $(0.1, 0.25, 0.5, 0.75, 0.9)$ quantiles (dashed black) of the estimated (orange) pdf are given by repeated sub-sampling from the estimated Lognormal CDFs and re-computation of (7.17). Panel (III) plots the log of breach sizes in excess of 10^3 ($n = 298$), that occurred at US organisations versus log org. size. The lines are linear quantile regressions of this data, where change points are automatically detected (additive quantile regression) when apparent in the data. The $q = 0.3$, 0.5, 0.7, and 0.9 quantile regressions are given. The estimated upward slopes, standard errors, and p-values, of these quantiles, for org. size below 10^{10}, are: $q = 0.3$: 0.66 (0.16) 0.00; $q = 0.5$: 0.40 (0.16) 0.013; $q = 0.7$: 0.23 (0.14) 0.02; $q = 0.9$: 0.18 (0.12) 0.39.

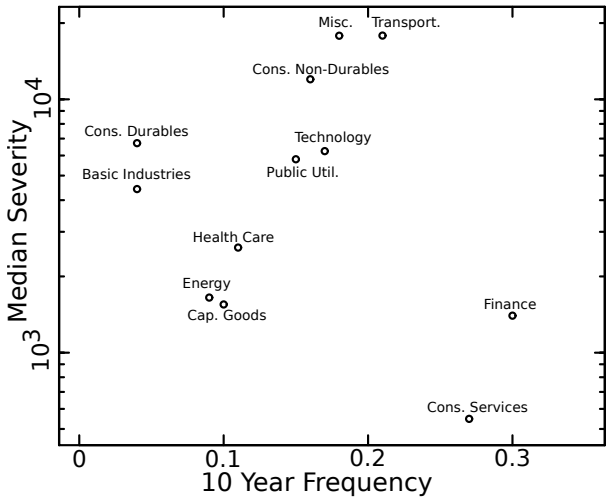

Figure 7.3: The median severity and 10 year frequency of breaches for the considered publicly traded firms, grouped into twelve sectors, as defined in [187].

Bibliography

[1] City population: Population statistics for countries, administrative areas, cities and agglomerations - interactive maps and charts. citypopulation.de. Accessed on 01-01-2015.

[2] Idaho National Labratory: Next Generation Nuclear Plant Probabilistic Risk Assessment White Paper, INL/EXT-11-21270. Idaho National Laboratory, Next Generation Nuclear Plant Project Idaho Falls, Idaho 83415, September 2011.

[3] List of major cyber attacks (wikipedia). http://en.wikipedia.org/wiki/List_of_cyber-attacks. Accessed: 2015-04-10.

[4] Nuclear neighbours. http://www.nature.com/news/2011/110421/full/472400a/box/3.html. Accessed: 2015-02-01.

[5] Nuclear safety chief calls for reform. http://www.nature.com/news/2011/110418/full/472274a.html. Accessed: 2015-02-01.

[6] UK Cabinet: The Cost of Cyber Crime. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60943/the-cost-of-cyber-crime-full-report.pdf. Accessed: 2015-09-01.

[7] European Commission: The different "generations" of nuclear technology. http://ec.europa.eu/research/energy/euratom/index_en.cfm?pg=fission&section=generation, 2015. accessed 01-12-2015.

[8] A. Acquisti, A. Friedman, and R. Telang. Is there a cost to privacy breaches? an event study. *ICIS 2006 Proceedings*, page 94, 2006.

[9] A. Acquisti and R. Gross. Predicting social security numbers from public data. *Proceedings of the National academy of sciences*, 106(27):10975–10980, 2009.

[10] C. Aggarwal. Outlier analysis. *Springer Science & Business Media*, 12, 2013.

[11] Y. Aït-Sahalia, J. Cacho-Diaz, and R. J. Laeven. Modeling financial contagion using mutually exciting jump processes. *NBER Working Paper No. 15850*, 2011.

[12] M. Aitkin and G. T. Wilson. Mixture models, outliers, and the em algorithm. *Technometrics*, 22(3):325–331, 1980.

[13] S. Albeverio, V. Jentsch, and H. Kantz. *Extreme events in nature and society*. Springer Science & Business Media, 2006.

[14] Allianz. Allianz Risk Barometer: Top Business Risks 2015. http://www.agcs.allianz.com/assets/PDFs/Reports/Allianz-Risk-Barometer-2015_EN.pdf. Accessed: 2015-09-01.

[15] L. Amaral, S. Buldyrev, S. Havlin, M. Salinger, and H. Stanley. Power law scaling for a system of interacting units with complex internal structure. *Physical Review Letters*, 80(7):1385–1388, 1998.

[16] J.-C. Anifrani, C. Le Floc'h, D. Sornette, and B. Souillard. Universal log-periodic correction to renormalization group scaling for rupture stress prediction from acoustic emissions. *Journal de Physique I*, 5(6):631–638, 1995.

[17] A. Ansar, B. Flyvbjerg, A. Budzier, and D. Lunn. Should We Build More Large Dams? The Actual Costs Of Mega-Dam Development. *Energy Policy*, 69:43–56, 2014.

[18] B. Appelbaum. As U.S. Agencies Put More Value on a Life, Businesses Fret. *The New York Times*, Feb 16, 2011.

[19] T. Aven. On how to deal with deep uncertainties in a risk assessment and management context. *Risk Analysis*, 33(12):2082–2091, 2013.

[20] T. Aven. On the meaning of a black swan in a risk context. *Safety science*, 57:44–51, 2013.

[21] E. Bacry, K. Dayri, and J. Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *Eur. Phys. J. B 85: 157*, 2012.

[22] E. Bacry and J.-F. Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, 2014.

[23] K. Balakrishnan. Exponential distribution: theory, methods and applications. *CRC press*, pages 228–230, 1996.

[24] U. Balasooriya and V. Gadag. Tests for upper outliers in the two-parameter exponential distribution. *Journal of statistical computation and simulation*, 50.3-4:249–259, 1994.

[25] V. Barnett and T. Lewis. Outliers in Statistical Data. 3rd ed. *John Wiley* , pages 285–293, 1994.

[26] A. D. Barnosky, E. A. Hadly, J. Bascompte, E. L. Berlow, J. H. Brown, M. Fortelius, W. M. Getz, J. Harte, A. Hastings, P. A. Marquet, et al. Approaching a state shift in earth/'s biosphere. *Nature*, 486(7401):52–58, 2012.

[27] J. Barry. The Great Influenza: The Epic Story of the Deadliest Plague in History. *New York, Penguin*, 2004.

[28] H. Bauke. Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B*, 58(2):167–173, 2007.

[29] L. Bauwens and N. Hautsch. Modelling Financial High Frequency Data Using Point Processes. In T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen, editors, *Handbook of Financial Time Series*, pages 953–979. Springer, May 2009.

[30] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

[31] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[32] J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4.510:126, 1998.

[33] J.-P. Bouchaud, J. Farmer, and F. Lillo. *Handbook of Financial Markets: Dynamics and Evolution*. North Holland, 2008.

[34] C. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models . *Journal of Econometrics*, 141(2):876–912, 2007.

[35] C. G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

[36] K. Brännäs and A. Hall. Estimation in integer-valued moving average models. *Applied Stochastic Models in Business and Industry*, 17(3):277–291, 2001.

[37] D. R. Brillinger. Time series, point processes, and hybrids. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 177–206, 1994.

[38] D. R. Brillinger. Statistical inference for stationary point processes. In *Selected Works of David Brillinger*, pages 499–543. Springer, 2012.

[39] P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer Science & Business Media, 2013.

[40] G. Brumfiel. Nuclear agency faces reform calls. *Nature*, 2011.

[41] D. Bundesbank. High-frequency trading and market implications. `https://www.bundesbank.de/Redaktion/EN/Downloads/Press/Presentations/2012_07_04_nagel_hft_und_martkimplikationen.pdf?__blob=publicationFile`. Accessed on 29-07-2015.

[42] P. Burgherr, P. Eckle, and S. Hirschberg. Comparative assessment of severe accident risks in the coal, oil and natural gas chains. *Reliability Engineering and System Safety*, 105:97–103, 2012.

[43] K. Campbell, L. A. Gordon, M. P. Loeb, and L. Zhou. The economic cost of publicly announced information security breaches: empirical evidence from the stock market. *Journal of Computer Security*, 11(3):431–448, 2003.

[44] P. Cauwels and D. Sornette. Quis pendit ipsa pretia: facebook valuation and diagnostic of a bubble based on nonlinear demographic dynamics. *Journal of portfolio management*, 38(2), 2012.

[45] H. L. d. S. Cavalcante, M. Oriá, D. Sornette, E. Ott, and D. J. Gauthier. Predictability and suppression of extreme events in a chaotic system. *Physical review letters*, 111(19):198701, 2013.

[46] R. E. Chandler. A spectral method for estimating parameters in rainfall models. *Bernoulli*, pages 301–322, 1997.

[47] V. Chavez-Demoulin, A. C. Davison, and A. J. McNeil. Estimating value-at-risk: a point process approach. *Quantitative Finance*, 5(2):227–234, 2005.

[48] V. Chavez-Demoulin and J. McGill. High-frequency financial data modeling using Hawkes processes . *Journal of Banking & Finance*, 36(12):3415–3426, 2012.

[49] D. Chernov and D. Sornette. *Man-made catastrophes and risk information concealment (25 case studies of major disasters and human fallibility)*. Springer, 2015.

[50] M. Chikkagoudar and S. Kunchur. Distributions of test statistics for multiple outliers in exponential samples. *Communications in StatisticsTheory and Methods*, 12:2127–2142, 1983.

[51] E. Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.

[52] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[53] M. Cohen. Epidemiology of drug resistance: implications for a postantimicrobial era. *Science*, 257(5073):1050–1055, 1992.

[54] S. Coles. An introduction to statistical modeling of extreme values. Vol. 208. *Springer*, 2001.

[55] R. Cont. Statistical Modeling of High Frequency Financial Data: Facts, Models and Challenges. *IEEE Signal Processing*, 28(5):16–25, 2011.

[56] A. Corral, F. Font, and J. Camacho. Noncharacteristic half-lives in radioactive decay. *Physical Review E*, 83(6):066103, 2011.

[57] C. Coughlin. Stuxnet virus attack: Russia warns of 'iranian chernobyl'. *The Telegraph (Jan 16, 2011)*.

[58] P. Cowpertwait. A Renewal Cluster Model for the Inter-Arrival Times of Rainfall Events. *International Journal of Climatology*, 21:49–61, 2001.

[59] P. Cowpertwait, P. O'Connell, A. Metcalfe, and J. Mawdsley. Stochastic point process modelling of rainfall. i. single-site fitting and validation. *Journal of Hydrology*, 175(1):17–46, 1996.

[60] D. Cox and V. Isham. *Point Processes*. CRC Press, 1980.

[61] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[62] D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes. Volume II: General theory and structure*, volume 1 of *Probability and Its Applications*. Springer Verlag, 2nd edition edition, 2003.

[63] J. Danielsson, L. de Haan, L. Peng, and C. G. de Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate analysis*, 76(2):226–248, 2001.

[64] A. Dassios and J.-W. Jang. Pricing of catastrophe reinsurance and derivatives using the cox process with shot noise intensity. *Finance and Stochastics*, 7(1):73–95, 2003.

[65] A. Dassios and J.-W. Jang. Kalman-bucy filtering for linear systems driven by the cox process with shot noise intensity and its application to the pricing of reinsurance contracts. *Journal of applied probability*, pages 93–107, 2005.

[66] A. Dassios and H. Zhao. A dynamic contagion process. *Advances in applied probability*, pages 814–846, 2011.

[67] A. Dassios, H. Zhao, et al. Exact simulation of hawkes process with exponentially decaying intensity. *Electronic Communications in Probability*, 18(62), 2013.

[68] Y.-A. de Montjoye, L. Radaelli, V. Singh, and A. Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

[69] W. Deemer and D. Votaw. Estimation of parameters of truncated or censored exponential distributions. *The Annals of Mathematical Statistics, 498-504*, 1955.

[70] A. Deluca and Á. Corral. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophysica*, 61(6):1351–1394, 2013.

[71] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journ. of the Royal Statistical Society. Serie. B (Methodological)*, pages 1–38, 1977.

[72] J. Dion, G. Gauthier, and A. Latour. Branching processes with immigration and integer-valued time series. *Serdica Mathematical Journal*, 21.2:123–136, 1995.

[73] W. Dixon. Analysis of Extreme Values. *The Annals of Mathematical Statistics*, pages 488–506, 1950.

[74] B. Dubrulle, F. Graner, and D. Sornette. Scale Invariance and Beyond. *Les Houches Workshop, March 10-14, 1997 (Centre de Physique des Houches), Springer*, 1998.

[75] J. Eeckhout. Gibrat's law for (all) cities. *American Economic Review*, 94:1429–1451, 2004.

[76] J. Eeckhout. Gibrat's law for (all) cities: Reply. *American Economic Review*, 99:1676–1683, 2009.

[77] A. Eleazar. *R. Abrahami Eleazaris Uraltes Chymisches Werck*. Crusius Erfurt, 1735.

[78] P. Embrechts and M. Kirchner. Hawkes graphs. *arXiv preprint arXiv:1601.01879*, 2016.

[79] P. Embrechts, C. Klüppelberg, and T. Mikosch. Modelling extremal events: for insurance and finance. Vol. 33. *Springer*, 1997.

[80] P. Embrechts, T. Liniger, and L. Lin. Multivariate Hawkes processes: an application to financial data . *Journal of Applied Probability*, 48:367–378, 2011.

[81] P. Embrechts, T. Liniger, L. Lin, et al. Multivariate hawkes processes: an application to financial data. *Journal of Applied Probability*, 48:367–378, 2011.

[82] P. Embrechts, S. I. Resnick, and G. Samorodnitsky. Extreme value theory as a risk management tool. *North American Actuarial Journal*, 3(2):30–41, 1999.

[83] V. Enciso-Mora, P. Neal, and T. Subba Rao. Efficient order selection algorithms for integervalued ARMA processes. *Journal of Time Series Analysis*, 30.1:1–18, 2009.

[84] K. Errais and L. Goldberg. Affine Point Processes and Portfolio Credit Risk. *SIAM Journal on Financial Mathematics 1*, 1:642–665, 2010.

[85] L. Escobar Rangel and F. Leveque. How Fukushima Dai-ichi core meltdown changed the probability of nuclear accidents? *Safety Science 64 90-98*, 2014.

[86] J. A. Feigenbaum and P. G. Freund. Discrete scale invariance in stock markets before crashes. *International Journal of Modern Physics B*, 10(27):3737–3745, 1996.

[87] S. Ferson, L. Ginzburg, and R. Akcakaya. Whereof one cannot speak: when input distributions are unknown. *Risk Analysis*, 1996.

[88] S. Ferson and L. R. Ginzburg. Different methods are needed to propagate ignorance and variability. *Reliability Engineering & System Safety*, 54(2):133–144, 1996.

[89] J. Ficenec. Cyber risk the most serious threat to business, says lloyd's chief. `http://www.telegraph.co.uk/finance/11516277/Cyber-risk-the-most-serious-threat-to-business-says-Lloyds-chief.html`. Accessed: 2015-09-01.

[90] V. Filimonov, D. Bicchetti, N. Maystre, and D. Sornette. Quantification of the High Level of Endogeneity and of Structural Regime Shifts in Commodity Prices. *The Journal of International Money and Finance*, 42(5):174–192, 2014.

[91] V. Filimonov and D. Sornette. Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108, 2012.

[92] V. Filimonov and D. Sornette. Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108, 2012.

[93] V. Filimonov and D. Sornette. Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data. *Quantitative Finance*, (in press), (http://ssrn.com/abstract=2371284), 2014.

[94] V. Filimonov and D. Sornette. Power law scaling and 'Dragon-Kings' in distributions of intraday financial drawdowns. *Chaos, Solitons & Fractals*, 74:27–45, 2015.

[95] B. Flyvbjerg, M. S. Holm, and S. Buhl. Underestimating Costs in Public Works Projects: Error or Lie? *Journal of the American Planning Association*, 68(3):279–295, 2002.

[96] B. Flyvbjerg, M. S. Holm, and S. Buhl. What Causes Cost Overrun in Transport Infrastructure Projects? *Transport Reviews*, 24(1):3–18, 2004.

[97] K. Fokianos. Count time series models. *Time Series – Methods and Applications*, 30:315–347, 2007.

[98] P. Fortune. Stock market crashes: what have we learned from October 1987. *New England Economic Review* , March/April:3–24, 1993.

[99] J. Franklin, V. Paxson, A. Perrig, and S. Savage. An inquiry into the nature and causes of the wealth of internet miscreants. In *Proceedings of the 14th ACM conference on Computer and communications security*, CCS '07, pages 375–388, New York, NY, USA, 2007. ACM.

[100] A. Garg, J. Curtis, and H. Halper. Quantifying the financial impact of it security breaches. *Information Management & Computer Security*, 11(2):74–83, 2003.

[101] K. M. Gatzlaff and K. A. McCullough. The effect of data breaches on shareholder wealth. *Risk Management and Insurance Review*, 13(1):61–83, 2010.

[102] R. Geller, Jackson, Y. Kagan, and F. Mulargia. Cannot earthquakes be predicted? *Science*, 278(5337):487–490, 1997.

[103] J. Gentle, W. Haerdle, and Y. Mori. *Handbook of computational statistics*. Springer, 2004.

[104] R. Gibrat. Les Inégalités Economiques; Applications aux inégalitiés des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel. *American Economic Review)*, Paris: Librarie du Recueil Sirey, 1931.

[105] M. I. Gomes and O. Oliveira. The bootstrap methodology in statistics of extremeschoice of the optimal sample fraction. *Extremes*, 4(4):331–358, 2001.

[106] D. Guha-Sapir, R. Below, and P. Hoyois. EM-DAT: The CRED/OFDA International Disaster Database, www.emdat.be, Université Catholique de Louvain, Brussels, Belgium. `www.emdat.be`. Accessed on 02-07-2015.

[107] H. Gupta. Gradually truncated log-normal in USA publicly traded firm size distribution. *Physica A: Statistical Mechanics and its Applications 375.2*, pages 643–650, 2007.

[108] M. Ha-Duong and V. Journe. Calculating nuclear accident probabilities from empirical frequencies. *Environment Systems and Decisions 34.2*, 2014.

[109] P. Hall, A. Welsh, et al. Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, 13(1):331–341, 1985.

[110] S. Hallegatte, A. Shah, C. Brown, R. Lempert, and S. Gill. Investment decision making under deep uncertainty–application to climate change. *World Bank Policy Research Working Paper*, (6193), 2012.

[111] P. Halpin. A Scalable EM Algorithm for Hawkes Processes . *New Developments in Quantitative Psychology*, pages 403–414, 2013.

[112] S. Hardiman, N. Bercot, and J. Bouchaud. Critical reflexivity in financial markets: a Hawkes process analysis. *European Journal of Physics B 86: 442*, 2013.

[113] E. Harrell and L. Langton. Victims of identity theft, 2012. *Washington DC: Bureau of Justice Statistics*, page 26, 2013.

[114] T. E. Harris. *The theory of branching processes*. Courier Corporation, 2002.

[115] D. Harte et al. Ptprocess: An r package for modelling marked point processes indexed by time. *Journal of Statistical Software*, 35(8):1–32, 2010.

[116] A. Hawkes. Point Spectra of Some Mutually Exciting Point Processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3):438–443, 1971.

[117] A. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[118] A. Hawkes and L. Adamopoulos. Cluster models for earthquakes-regional comparisons. *Bull. Int. Statist. Inst*, 45(3):454–461, 1973.

[119] A. Hawkes and D. Oakes. A Cluster Process Representation of a Self Exciting Process. *J.Appl.Prob 11*, pages 497–503, 1974.

[120] D. Hawkins. Identification of outliers. Vol. 11. *Chapman and Hall*, 1980.

[121] A. Helmstetter and D. Sornette. Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *Journal of Geophysical Research: Solid Earth (1978–2012)*, 107(B10):ESE–10, 2002.

[122] B. M. Hill et al. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 3(5):1163–1174, 1975.

[123] S. Hirschberg, C. Bauer, P. Burgherr, E. Cazzoli, T. Heck, M. Spada, and K. Treyer. Health effects of technologies for power generation: Contributions from normal operation, severe accidents and terrorist threat. *Reliability Engineering and System Safety*, 2015.

[124] S. Hirschberg, P. Burgherr, G. Spiekerman, and R. Dones. Severe accidents in the energy sector: comparative perspective. *Journal of Hazardous Materials*, 111(1):57–65, 2004.

[125] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[126] M. Hofert and M. V. Wüthrich. Statistical Review of Nuclear Power Accidents. *Asia-Pacific Journal of Risk and Insurance*, 7(1):1–18, 2013.

[127] K. Hsü. Nuclear Risk Evaluation. *Nature 328, 22*, 1987.

[128] Y. Huang, G. Ouillon, H. Saleur, and D. Sornette. Spontaneous generation of discrete scale invariance in growth models. *Physical Review E*, 55(6):6433, 1997.

[129] IAEA. Development and Application of Level 1 Probabilistic Safety Assessment for Nuclear Power Plants. IAEA Safety Standards No. SSG-3, Specific Safety Guide, INTERNATIONAL ATOMIC ENERGY AGENCY, VIENNA, April 2010.

[130] IAEA. IAEA Safety Standards for protecting people and the environment, Development and Application of Level 2 Probabilistic Safety Assessment for Nuclear Power Plants. Specific Safety Guide, No. SSG-4, INTERNATIONAL ATOMIC ENERGY AGENCY, VIENNA, 2010.

[131] IAEA. IAEA Safety Standards for protecting people and the environment, Development and Application of Level 2 Probabilistic Safety Assessment for Nuclear Power Plants. Specific Safety Guide, No. SSG-4, INTERNATIONAL ATOMIC ENERGY AGENCY, VIENNA, 2010.

[132] IAEA. Measures to strengthen international cooperation in nuclear, radiation and waste safety. `https://www.iaea.org/About/Policy/GC/GC40/GC40InfDocuments/English/gc40inf-5_en.pdf`. Specific Safety Guide, No 50-P-12, INTERNATIONAL ATOMIC ENERGY AGENCY,VIENNA, 1996.

[133] IAEA. Procedures for Conducting Probabilistic Safety Assessments of Nuclear Power Plants (Level 3), Off-site consequences and estimation of risks to the public. Specific Safety Guide, No 50-P-12, INTERNATIONAL ATOMIC ENERGY AGENCY, VIENNA, 1996.

[134] IAEA. Nuclear power reactors in the world 2010. *Reference data series no. 2. 2010 Edition*, 2010.

[135] IAEA. INES User Manual. `http://www-pub.iaea.org/MTCD/Publications/PDF/INES2013web.pdf`, 2013. accessed 01-12-2015.

[136] B. Iglewicz and J. Martinez. Outlier detection using robust measures of scale. *Journal of Statistical Computation and Simulation* , 15.4:285–293, 1982.

[137] IRSN. BEMUSE project. `http://www.irsn.fr/EN/Research/Research-organisation/Research-programmes/BEMUSE/Pages/BEMUSE-project-4481.aspx`, 2015. accessed 01-12-2015.

[138] J. Janczura and R. Weron. Black swans or dragon-kings? a simple test for deviations from the power law. *The European Physical Journal Special Topics*, 205(1):79–93, 2012.

[139] A. Johansen and D. Sornette. Stock Market Crashes are outliers. *The European Physical Journal B-Condensed Matter and Complex Systems*, 1.2:141–143, 1998.

[140] A. Johansen and D. Sornette. Critical ruptures. *The European Physical Journal B-Condensed Matter and Complex Systems*, 18(1):163–181, 2000.

[141] A. Johansen and D. Sornette. Large stock market price drawdowns are outliers. *Journal of Risk*, 4:69–110, 2002.

[142] A. Johansen, D. Sornette, et al. Shocks, crashes and bubbles in financial markets. *Brussels Economic Review (Cahiers economiques de Bruxelles)*, 53(2):201–253, 2010.

[143] M. Kacperczyk and S. P. When Safe Proved Risky: Commercial Paper during the Financial Crisis of 2007-2009. *Journal of Economic Perspectives*, 24(1):29–50, 2010.

[144] M. Kandlikar, J. Risbey, and S. Dessai. Representing and communicating deep uncertainty in climate-change assessments. *Comptes Rendus Geoscience*, 337(4):443–455, 2005.

[145] C. W. Karvetski and J. H. Lambert. Evaluating deep uncertainties in strategic priority-setting with an application to facility energy investments. *Systems Engineering*, 15(4):483–493, 2012.

[146] A. Y. Khinchin. Sequences of chance events without after-effects. *Theory of Probability & Its Applications*, 1(1):1–15, 1956.

[147] A. Kimber. Tests for many outliers in an exponential sample. *Applied Statistics*, pages 263–271, 1982.

[148] C. Kindleberger. *Manias, Panics, and Crashes: A History of Financial Crises)*. (Wiley Investment Classics), Wiley, 4th edition, 2000.

[149] M. Kirchner. An estimation procedure for the hawkes process. *arXiv preprint arXiv:1509.02017*, 2015.

[150] M. Kirchner. Hawkes and inar (∞) processes. *arXiv preprint arXiv:1509.02007*, 2015.

[151] C. Klüppelberg and T. Mikosch. Explosive poisson shot noise processes with applications to risk reserves. *Bernoulli*, pages 125–147, 1995.

[152] F. H. Knight. *Risk, uncertainty and profit*. Courier Corporation, 2012.

[153] R. Koenker. Quantile regression. No. 38. . *Cambridge university press*, 2005.

[154] C. Kooperberg and C. J. Stone. A study of logspline density estimation. *Computational Statistics & Data Analysis*, 12(3):327–347, 1991.

[155] T. Kovalenko and D. Sornette. Dynamical Diagnosis and Solutions for Resilient Natural and Social Systems. *Planet Risk (Davos, Global Risk Forum (GRF) Davos)*, 1(1):7–33, 2013.

[156] W. Kröger and D. Sornette. Reflections on Limitations of Current PSA Methodology. *ANS PSA 2013 International Topical Meeting on Probabilistic Safety Assessment and Analysis, Columbia, South Carolina, USA, September 22-26*, 2013.

[157] J. Laaksonen. Thoughts in the aftermath of accident at the Fukushima Daiichi NPP. *PSAM11 - ESREL 2012 Helsinki, Finland June 26, 2012*, 2012.

[158] J. Laherrère and D. Sornette. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2.4:525–539, 1998.

[159] J. C. Lee and N. J. McCormick. *Risk and safety analysis of nuclear systems*. John Wiley & Sons, 2012.

[160] A. Lesne and M. Laguës. Scale Invariance: From Phase Transitions to Turbulence. *Springer (2012 edition)*, 2011.

[161] N. Leveson. *MIT, Technical and Managerial Factors in the NASA Challenger and Columbia Losses: Looking Forward to the Future, published within Kleinman, Cloud-Hansen, Matta, and Handelsman, Controveries in Science and Technology Volume 2*. Mary Ann Liebert Press, 2008.

[162] M. Levy. Gibrat's law for (all) cities: A comment. *American Economic Review*, 99:1672–1675, 2009.

[163] E. Lewis and G. Mohler. A Nonparametric EM algorithm for Multiscale Hawkes Processes. *J. Nonparametric Statistics*, 2011.

[164] E. Lewis and G. Mohler. A Nonparametric EM algorithm for Multiscale Hawkes Processes. *Preprint*, pages 1–16, May 2011.

[165] E. Lewis, G. Mohler, P. Brantingham, and A. Bertozzi. Self-exciting point process models of civilian deaths in Iraq. *Security Journal*, 25:244–264, 2012.

[166] P. A. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

[167] T. Lewis and N. Fieller. A recursive algorithm for null distributions for outliers: I Gamma Samples. *Technometrics*, 21:371–376, 1979.

[168] J. Likeš. Distribution of Dixon's statistics in the case of an exponential population. *Metrika* , 11(1):46–54, 1967.

[169] C. Lin and N. Balakrishnan. Exact computation of the null distribution of a test for multiple outliers in an exponential sample. *Computational Statistics & Data Analysis* , 53.9:3281–3290, 2009.

[170] C. Lin and N. Balakrishnan. Tests for Multiple Outliers in an Exponential Sample. *Communications in Statistics – Simulation and Computation* , 43.4:706–722, 2014.

[171] I. Linkov, E. Anklam, Z. A. Collier, D. DiMase, and O. Renn. Risk-based standards: integrating top–down and bottom–up approaches. *Environment Systems and Decisions*, 34(1):134–137, 2014.

[172] D. Lochbaum. Nuclear Plants Risk Studies: Failing the grade. *Union of concerned scientists report*, 2000.

[173] L. M. B. Cabral and J. Mata. On the Evolution of the Firm Size Distribution: Facts and Theory. *The American Economic Review*, 2003.

[174] T. Maillart and D. Sornette. Heavy-tailed distribution of cyber-risks. *Eur. Phys. J. B*, 75(3):357–364, 2010.

[175] Y. Malevergne, V. Pisarenko, and D. Sornette. Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Physical Review E*, 83:036111, 2011.

[176] Y. Malevergne and D. Sornette. *Extreme financial risks: From dependence to risk management.* Springer Science & Business Media, 2006.

[177] B. Mandelbrot and R. Hudson. The Misbehavior of Markets: A fractal view of financial turbulence. *Basic Books*, 2014.

[178] D. Marsan and O. Lengline. Extending Earthquakes' Reach Through Cascading. *Science*, 319(5866):1076–1079, Feb. 2008.

[179] F. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46:68–78, 1951.

[180] McAfee. McAfee Unsecured Economies Report. 2008.

[181] E. McKenzie. Discrete variate time series. *Stochastic Processes: Modelling and Simulation*, page 573606, 2003.

[182] T. Mikosch. Non-Life Insurance Mathematics. 2nd printing. . *Springer*, 2006.

[183] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[184] J. Møller and J. Rasmussen. Perfect simulation of Hawkes processes. *Advances in applied probability*, 37(3):629–646, 2005.

[185] P. Morzywolek. Non-parametric methods for estimation of hawkes process for high-frequency financial data. *ETH Zürich Master Thesis*, 2015.

[186] A. Mosleh. Delivering on the Promise: PRA, Real Decisions, and Real Events. *PSAM11 - ESREL 2012 Helsinki, Finland June 29, 2012, Plenary Talk PSAM11-ESREL2012*, 2012.

[187] Nasdaq. Company list (nasdaq, nyse, & amex). `http://www.nasdaq.com/screening/company-list.aspx`. Accessed: 2014-10-01.

[188] P. Neal and T. Subba Rao. MCMC for IntegerValued ARMA processes. *Journal of Time Series Analysis*, (28.1):92–110, 2007.

[189] M. E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[190] D. Oakes. The Markovian Self-Exciting Process. *Applied Probability Trust*, 12(1):69–77, 1975.

[191] Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261, 1978.

[192] Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, Jan. 1981.

[193] Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

[194] Y. Ogata. A Prospect of Earthquake Prediction Research. *Statistical Science*, 28(4):521–541, 2013.

[195] E. Ohlsson and B. Johansson. Non-life insurance pricing with generalized linear models. *Springer Science & Business Media*, 2010.

[196] OSF. Open security foundation data loss database. `http://datalossdb.org/`. Accessed: 2015-04-10.

[197] T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, Dec. 1979.

[198] F. Papangelou. Integrability of expected increments of point processes and a related random change of scale. *Transactions of the American Mathematical Society*, 165:483–506, 1972.

[199] C. Perrow. *Normal Accidents: Living with High-Risk Technologies*. Princeton University Press, Princeton, New Jersey, 2nd edition, 1999.

[200] V. Pisarenko and M. V. Rodkin. *Heavy-tailed distributions in disaster analysis*, volume 30. Springer Science & Business Media, 2010.

[201] V. Pisarenko and D. Sornette. Robust statistical tests of Dragon-Kings beyond power law distributions. *The European Physical Journal-Special Topics*, 205(1):95–115, 2011.

[202] Ponemon Institute. 2014 cost of data breach study: United states. 2014.

[203] C. Potter. A History of Influenza. *J Appl Microbiol.*, 91(4):572–579, 2006.

[204] PRCH. Privacy rights clearing house. `http://www.privacyrights.org/`. Accessed: 2015-04-10.

[205] M. Prokešová and E. B. V. Jensen. Asymptotic palm likelihood theory for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 65(2):387–412, 2013.

[206] PWC. PwC 18th Annual Global CEO Survey. `http://www.pwc.com/gx/en/ceo-agenda/ceo-survey.html`. Accessed: 2015-09-01.

[207] Rand. Markets for cybercrime tools and stolen data: Hackers bazaar. `http://www.rand.org/content/dam/rand/pubs/research_reports/RR600/RR610/RAND_RR610.pdf`. Accessed: 2015-01-01.

[208] K. D. Rao, H. Kushwaha, A. K. Verma, and A. Srividya. Quantification of epistemic and aleatory uncertainties in level-1 probabilistic safety assessment studies. *Reliability Engineering & System Safety*, 92(7):947–956, 2007.

[209] T. S. Rao and R. Chandler. A frequency domain approach for estimating parameters in point process models. In *Athens Conference on Applied Probability and Time Series Analysis*, pages 392–405. Springer, 1996.

[210] R. Redner and H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26.2:195–239, 1984.

[211] A. Rényi. On the theory of order statistics. *Acta Mathematica Hungarica* , 4.3:191–231, 1953.

[212] T. Reynkens, J. Beirlant, J. De Spiegeleer, K. Herrmann, and W. Schoutens. Hunting for Black Swans in the European banking sector using extreme value analysis. *9th International EVA conference. Ann Arbor, MI*, pages 14–19, 2015.

[213] I. Rodriguez-Iturbe, D. Cox, and V. Isham. Some models for rainfall based on stochastic point processes. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 410, pages 269–288. The Royal Society, 1987.

[214] S. Rogers. Nuclear power plant accidents: listed and ranked since 1952. `http://www.theguardian.com/news/datablog/2011/mar/14/nuclear-power-plant-accidents-list-rank`. Accessed: 2015-02-01.

[215] B. Rosner. On the detection of many outliers. *Technometrics*, 17.2:221–227, 1975.

[216] A. Saichev, Y. Malevergne, and D. Sornette. Theory of Zipf's law and beyond. *Lecture Notes in Economics and Mathematical Systems*, 632:Springer, ISBN: 978–3–642–02945–5, 2009.

[217] A. Saichev and D. Sornette. Super-linear scaling of offsprings at criticality in branching processes. *Physical Review E*, 89:012104, 2014.

[218] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. Optimization with em and expectation-conjugate-gradient. In *ICML*, pages 672–679, 2003.

[219] A. Salim. Extensions of the Bartlett-Lewis Model for Rainfall Processes. *Statistical Modelling*, 3:79–98, 2003.

[220] C. G. Sammis and D. Sornette. Positive feedback, memory, and the predictability of earthquakes. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2501–2508, 2002.

[221] D. Sanger. *Confront and Conceal: Obama's Secret Wars and Surprising Use of American Power*. Crown Publishing Group, 2012.

[222] M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. Van De Koppel, I. A. Van De Leemput, S. A. Levin, E. H. Van Nes, et al. Anticipating critical transitions. *science*, 338(6105):344–348, 2012.

[223] E. Schlosser. *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*. Penguin Books; Reprint edition, 2014.

[224] C. Schluter and M. Trede. Identifying multiple outliers in heavy-tailed distributions with an application to market crashes. *Journal of Empirical Finance*, 15(4):700–713, 2008.

[225] M. Schneider and A. Froggatt. World Nuclear Industry status Report 2013. *Mycle Schneider Consulting*, 2013.

[226] C. Scholz. *The Mechanics of Earthquakes and Faulting*. Cambridge University Press; 2 edition, 2002.

[227] A. Sengör. Evaluating Nuclear Accidents. *Nature 335, 391*, 1987.

[228] A. Sengör. Predicting the threat of nuclear disasters. *Nature 335, 391*, 1987.

[229] B. E. Shaw, J. M. Carlson, and J. S. Langer. Patterns of seismic activity preceding large earthquakes. *Journal of Geophysical Resarch: Solid Earth*, 97(B1):479–488, 1992.

[230] P. Sieczka, D. Sornette, and J. Holyst. The Lehman Brothers Effect and Bankruptcy Cascades. *European Physical Journal B*, 82(3-4):257–269, 2011.

[231] M. E. Silva and V. L. Oliveira. Difference equations for the higher order moments and cumulants of the inar (p) model. *Journal of Time Series Analysis*, 26(1):17–36, 2005.

[232] B. Silverman. Density Estimation for Statistics and Data Analysis. *Chapman and Hall*, 1986.

[233] P. Simon. *Too Big to Ignore: The Business Case for Big Data*. John Wiley & Sons, 2013.

[234] L. Simonsen, P. Spreeuwenberg, R. Lustig, R. Taylor, D. Fleming, M. Kroneman, M. Van Kerkhove, A. Mounts, and W. Paget. (the GLaMOR Collaborating Teams), Global mortality estimates for the 2009 influenza pandemic from the GLaMOR project: a modeling study. *PLOS Medicine*, 10(11):e1001558, 2013.

[235] G. Sinanaj. News media sentiment of data breaches. 2014.

[236] D. Smythe. An objective nuclear accident magnitude scale for quantification of severe and catastrophic events. *Physics Today: Points of View*, 2011.

[237] K. Soramäki, M. L. Bech, J. Arnold, R. J. Glass, and W. E. Beyeler. The topology of interbank payment flows. *Physica A*, 379:317–333, 2007.

[238] D. Sornette. Discrete-scale invariance and complex dimensions. *Physics reports*, 297(5):239–270, 1998.

[239] D. Sornette. Linear stochastic dynamics with nonlinear fractal properties. *Physica A: Statistical Mechanics and its Applications 250.1*, pages 295–314, 1998.

[240] D. Sornette. Multiplicative processes and power law. *Physical Review E 57.4*, 1998.

[241] D. Sornette. Predictability of catastrophic events: Material rupture, earthquakes, turbulence, financial crashes, and human birth. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2522–2529, 2002.

[242] D. Sornette. *Why Stock Markets Crash (Critical Events in Complex Financial Systems)*. Princeton University Press, 2003.

[243] D. Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer Science & Business Media, 2006.

[244] D. Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer Science & Business Media, 2006.

[245] D. Sornette. Endogenous versus exogenous origins of crises. In *Extreme events in nature and society*, pages 95–119. Springer, 2006.

[246] D. Sornette. Dragon-kings, black swans and the prediction of crises. *Swiss Finance Institute Research Paper*, (09-36), 2009.

[247] D. Sornette. Dragon-Kings, Black Swans and the Prediction of Crises. *International Journal of Terraspace Science and Engineering)*, 2(1):1–18, 2009.

[248] D. Sornette. A civil super-apollo project in nuclear research for a safer and prosperous world. *Energy Research & Social Science*, 8:60–65, 2015.

[249] D. Sornette and P. Cauwels. 1980–2008: The illusion of the perpetual money machine and what it bodes for the future. *Risks*, 2(2):103–131, 2014.

[250] D. Sornette and P. Cauwels. Managing risk in a creepy world. *Journal of Risk Management in Financial Institutions*, 8(1):83–108, 2015.

[251] D. Sornette and R. Cont. Convergent multiplicative processes repelled from zero: power laws and truncated power laws. *Journal de Physique I 7.3*, pages 431–444, 1997.

[252] D. Sornette and A. Johansen. A Hierarchical Model of Financial Crashes. *Physica A*, 261:581–598, 1998.

[253] D. Sornette, A. Johansen, and J.-P. Bouchaud. Stock market crashes, precursors and replicas. *Journal de Physique I*, 6(1):167–175, 1996.

[254] D. Sornette, A. Johansen, and J.-P. Bouchaud. Stock market crashes, Precursors and Replicas. *J.Phys.I France*, 6(1):167–175, 1996.

[255] D. Sornette and T. Kovalenko. Dynamical diagnosis and solutions for resilient natural and social systems. *Planet@ Risk*, 1(1), 2013.

[256] D. Sornette, T. Maillart, and W. Kröger. Exploring the limits of safety analysis in complex technological systems. *International Journal of Disaster Risk Reduction*, 9:59–66, 2013.

[257] D. Sornette, T. Maillart, and W. Kröger. Exploring the limits of safety analysis in complex technological systems. *International Journal of Disaster Risk Reduction*, 6:59–66, Dec. 2013.

[258] D. Sornette, P. Miltenberger, and C. Vanneste. Statistical physics of fault patterns self-organized by repeated earthquakes. *Pure and Applied Geophysics*, 142(3-4):491–527, 1994.

[259] D. Sornette and G. Ouillon. Dragon-kings: mechanisms, statistical methods and empirical evidence. *The European Physical Journal-Special Topics*, 205(1):1–26, 2012.

[260] D. Sornette and S. Utkin. Limits of declustering methods for disentangling exogenous from endogenous events in time series with foreshocks, main shocks, and aftershocks. *Physical Review E*, 79(6):061110, June 2009.

[261] D. Sornette, R. Woodard, W. Yan, and W.-X. Zhou. Clarifications to questions and criticisms on the johansen–ledoit–sornette financial bubble model. *Physica A: Statistical Mechanics and its Applications*, 392(19):4417–4428, 2013.

[262] D. Sornette, R. Woodard, and W.-X. Zhou. The 2006–2008 oil bubble: Evidence of speculation, and prediction. *Physica A: Statistical Mechanics and its Applications*, 388(8):1571–1576, 2009.

[263] B. Sovacool. The Costs of Failure: A Preliminary Assessment of Major Energy Accidents, 1907 to 2007. *Energy Policy*, 35(5):1802–1820, 2008.

[264] B. Sovacool and M. Dworkin. Global Energy Justice: Problems, Principles, and Practices. *Cambridge: Cambridge University Press*, 2014.

[265] B. Sovacool, A. Gilbert, and D. Nugent. Risk, Innovation, Electricity Infrastructure and Construction Cost Overruns: Testing Six Hypotheses. *Energy*, 74:906–917, 2014.

[266] B. Sovacool, D. Nugent, and A. Gilbert. An International Comparative Assessment of Construction Cost Overruns for Electricity Infrastructure. *Energy Research & Social Science*, 3:152–160, 2014.

[267] H. Stephens and C. Bonini. The Size Distribution of Business Firms. *The American Economic Review*, 1958.

[268] M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association 69.347*, 1974.

[269] S. H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Westview press, 2014.

[270] L. Subramaniam and N. Kumar. Multiple outlier test for upper outliers in an exponential sample. *Journal of Applied Statistics*, 39.6:1323–1330, 2012.

[271] N. Taleb. *The black swan: the impact of the highly improbable*. Random House, 2010.

[272] U. Tanaka, Y. Ogata, and D. Stoyan. Parameter estimation and model selection for neyman-scott point processes. *Biometrical Journal*, 50(1):43–57, 2008.

[273] I. Toke. Market making in an order book model and its impact on the spread. *Econophysics of Order-Driven Markets, Springer Verlag*, pages 49–64, 2011.

[274] B. Turlach. Bandwidth selection in kernel density estimation: A review. *CORE and Institut de Statistique 19.4*, pages 1–33, 1993.

[275] R. A. Tybout. Pricing pollution and other negative externalities. *The Bell Journal of Economics and Management Science*, pages 252–266, 1972.

[276] UNAIDS. MDG 6: 15 years, 15 lessons of hope from the AIDS reponse. `http://www.unaids.org/sites/default/files/media_asset/20150714_FS_MDG6_Report_en.pdf`. Accessed on 29-07-2015.

[277] UNAIDS. Report on the global AIDS epidemic 2012. http://www.unaids.org/sites/default/files/en/media/unaids/contentassets/documents/epidemiology/2012/gr2012/20121120_UNAIDS_Global_Report_2012_with_annexes_en.pdf. Accessed on 29-07-2015.

[278] T. Utsu and Y. Ogata. The centenary of the omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth*, 43(1):1–33, 1995.

[279] A. Veen and F. Schoenberg. Estimation of space-time branching process models in seismology using an EM-type algorithm. *J. Amer. Statist. Assoc., 103*, pages 614–624, 2008.

[280] I. Verdinelli and L. Wasserman. Bayesian analysis of outlier problems using the gibbs sampler. *Statistics and Computing*, 1(2):105–117, 1991.

[281] D. Vere-Jones. Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–62, 1970.

[282] Verizon. 2014 Data Breach Investigations Report. 2014.

[283] Y. Virkar, A. Clauset, et al. Power-law distributions in binned empirical data. *The Annals of Applied Statistics*, 8(1):89–119, 2014.

[284] WEF. World Economic Forum: Global Risks 2015. http://www3.weforum.org/docs/WEF_Global_Risks_2015_Report15.pdf. Accessed: 2015-09-01.

[285] C. H. Weiß. Serial dependence and regression of poisson inarma models. *Journal of Statistical Planning and Inference*, 138(10):2975–2990, 2008.

[286] C. H. Weiß. Thinning operations for modeling time series of countsa survey. *AStA Advances in Statistical Analysis*, 92(3):319–341, 2008.

[287] S. Wheatley, V. Filimonov, and D. Sornette. The hawkes process with renewal immigration & its estimation with an em algorithm. *Computational Statistics & Data Analysis*, 94:120–135, 2016.

[288] S. Wheatley, P. Kyriakis, D. Sornette, and B. Sovacool. Nuclear events database. https://tasmania.ethz.ch/index.php/Nuclear_events_database.

[289] S. Wheatley, T. Maillart, and D. Sornette. The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B*, 89(1):1–12, 2016.

[290] S. Wheatley and D. Sornette. Multiple outlier detection in samples with exponential & pareto tails: Redeeming the inward approach & detecting dragon kings. *Submitted to Journal of Applied Statistics. arXiv preprint arXiv:1507.08689*, 2015.

[291] S. Wheatley, B. Sovacool, and D. Sornette. Of Disasters and Dragon Kings: A Statistical Analysis of Nuclear Power Incidents & Accidents. *Risk Analysis (submitted 7 April 2015)*.

[292] S. Wheatley, B. Sovacool, and D. Sornette. Of disasters and dragon kings: A statistical analysis of nuclear power incidents & accidents. *To appear in Risk Analysis*, 2016.

[293] S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[294] WNA. Economics of Nuclear Power. www.world-nuclear.org/info/Economic-Aspects/Economics-of-Nuclear-Power.

[295] C. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics 11.1*, pages 95–103, 1983.

[296] M. Wüthrich. Non-Life Insurance: Mathematics & Statistics. *SSRN Manuscript 2319328.*, 2014.

[297] L. Xu and M. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation*, 8:129–151, 1995.

[298] S. L. Zabell. Predicting the unpredictable. *Synthese*, 90(2):205–232, 1992.

Curriculum Vitae

CV

Name: WHEATLEY, Spencer
Address: 71 Herbstweg, 8050 Zürich
Email: spencerwheatley@gmail.com
Mobile: +41 78 909 8488
Birth: 15.09.1988
Nationality: Canadian with Swiss Residence Permit B
Position: PhD student at ETH Zürich

SUMMARY

Statistical study of extreme risks, and point process models in the group of Prof. Didier Sornette

Comprehensive education focused on statistics, risk management, and engineering sciences

Advanced skills in statistical mathematics, modeling, and computing (primarily using R)

Two years of professional experience spanning from engineering to management consulting

EDUCATION

ETH Zürich 10.2013 – 04.2016
PhD Science supervised by Prof. Didier Sornette
Thesis: Extending the Hawkes process, a general outlier test, & case studies in extreme risk

ETH Zürich 09.2011 – 08.2013
Master of Science ETH in Statistics. GPA: 5.7/6
Thesis: Non-parametric estimation of the Hawkes process, and application to financial data

University of Waterloo 09.2006 – 08.2011
BASc Honours Systems Design Engineering, Statistics Option (Co-operative)
GPA: 9/10. Top overall graduate award, Fall 2011

PRIOR RESEARCH EXPERIENCE

ETH Zürich, D-MTEC 09.2012 – 02.2013
Graduate Research Assistant under Prof. Didier Sornette Zürich, Switzerland
Hawkes process modeling of high frequency financial data

University of Waterloo, Dept. of Statistics and Actuarial Science 04.2011 – 08.2011
NSERC Undergraduate Research Assistant Waterloo, Canada
Modeling and analysis of network graphs; R programming and seminar presentations

PROFESSIONAL EXPERIENCE

Deloitte Inc. 01.2009 – 09.2009, 05.2010 – 09.2010
Analyst, Technology Strategy Consulting Toronto, Canada
Provided business-IT consulting services to executive-level clients
Developed business cases, IT system architectures, and business process designs for IT transformations in financial institutions

IBM, Cognos Business Intelligence 05.2008 – 09.2008
Technical Analyst Ottawa, Canada
Provided technical support to high profile clients

General Dynamics Canada 08.2007 – 12.2007
Human Factors Engineering / Interface Design Ottawa, Canada
Designed modules of the Maritime Helicopter mission system user interface
Managed helicopter software and usability requirements (Telelogic DOORS)

Vancouver Pile Driving Ltd. 01.2007 – 04.2007
Assistant Project Engineering Vancouver, Canada
Logistics technician for a variety of marine civil engineering construction projects

AWARDS

University of Waterloo, Top Overall Graduate, Fall 2011

Canadian Scholarship Trust Fund Kenneth Le M. Carter Graduate Award (2011)

Top Graduate: Sandford Fleming Award for Co-operative Proficiency (2011)

NSERC Undergraduate Student Research Award (2011)

Ontario International Exchange Scholarship (2009)

Waterloo President's Entrance Scholarship (2006)

ACADEMIC PUBLICATIONS

Wheatley, S. and Sornette, D. "Multiple Outlier Detection in Samples with Exponential & Pareto Tails: Redeeming the Inward Approach & Detecting Dragon Kings." arXiv:1507.08689 (2015).

Wheatley, S., Maillart, T., and Sornette, D. "The Extreme Risk of Personal Data Breaches & The Erosion of Privacy." European Physical Journal B 1.89 (2016): 1-12

Wheatley, S., Sovacool, B., and Sornette, D. "Of Disasters and Dragon Kings: A Statistical Analysis of Nuclear Power Incidents & Accidents." arXiv:1504.02380 (2015).

Wheatley, S., Benjamin Sovacool, and Didier Sornette. “Reassessing the safety of nuclear power.” Energy Research & Social Science (2016)

Wheatley, S., Filimonov, V., and Sornette, D., “Estimation of the Hawkes Process With Renewal Immigration Using the EM Algorithm”. Computational Statistics & Data Analysis 94 (2016):120-135.

Filimonov, V., Wheatley, S., and Sornette, D. "Effective measure of endogeneity for the Autoregressive Conditional Duration point processes via mapping to the self-excited Hawkes process." Communications in Nonlinear Science and Numerical Simulation 22.1 (2015): 23-37.

ACADEMIC PRESENTATIONS

Statistical Analysis of the Risk of Nuclear Accidents: New data and extreme heavy tailed distributions, ESREL conference, ETH Zürich, September 2015.

A general outlier test and case studies in extremes, Large fluctuations and extreme events summer school, TU Dresden, October 2015.

Hawkes process with renewal immigration and application to high frequency price fluctuations, Rmetrics summer school, Meielisalp, Switzerland, 2013