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

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.