Lunds tekniska högskola
Mathematical Statistics

Modelling News Sentiment Flow Using Spatial Hawkes Processes: Dependencies Between Topics and Countries

Carl Brishammar

Master Thesis
June 19, 2018

Abstract

This Master Thesis proposes to use a Spatial Hawkes Process to model the news flow in the world. The spatial part is divided in both geographical areas as well as different topics. Therefore, in the Hawkes model every news article corresponds to a point in space and time. In a certain region and topic the intensity of the released articles will be modelled with the Hawkes Process. This can be of interest for various applications depending on the topic and region chosen. The data in this project comes from the company RavenPack which has labelled every news article with a topic and a region. The area specifically examined in this report will be the relations between the different areas in the news flow. A comparison will also be done between some different spatial divisions to see if different behaviour can be captured with a more complex model, with more regions and topics.

The model is compared to a Poisson Process model. It seems that the Hawkes model works better than the Poisson Process to model the intensity of the different parts of the news flow in all cases. The results also indicates that a very flexible model will be able to capture more cases that are known from history. The complex model sees connections that increase the intensity during hectic times in the recent past news flow, for example during Brexit and the Arab Spring. However it seems that with more regions the model is prone to overfit and a simpler model may be preferable for out of sample uses.

Popular Science Summary

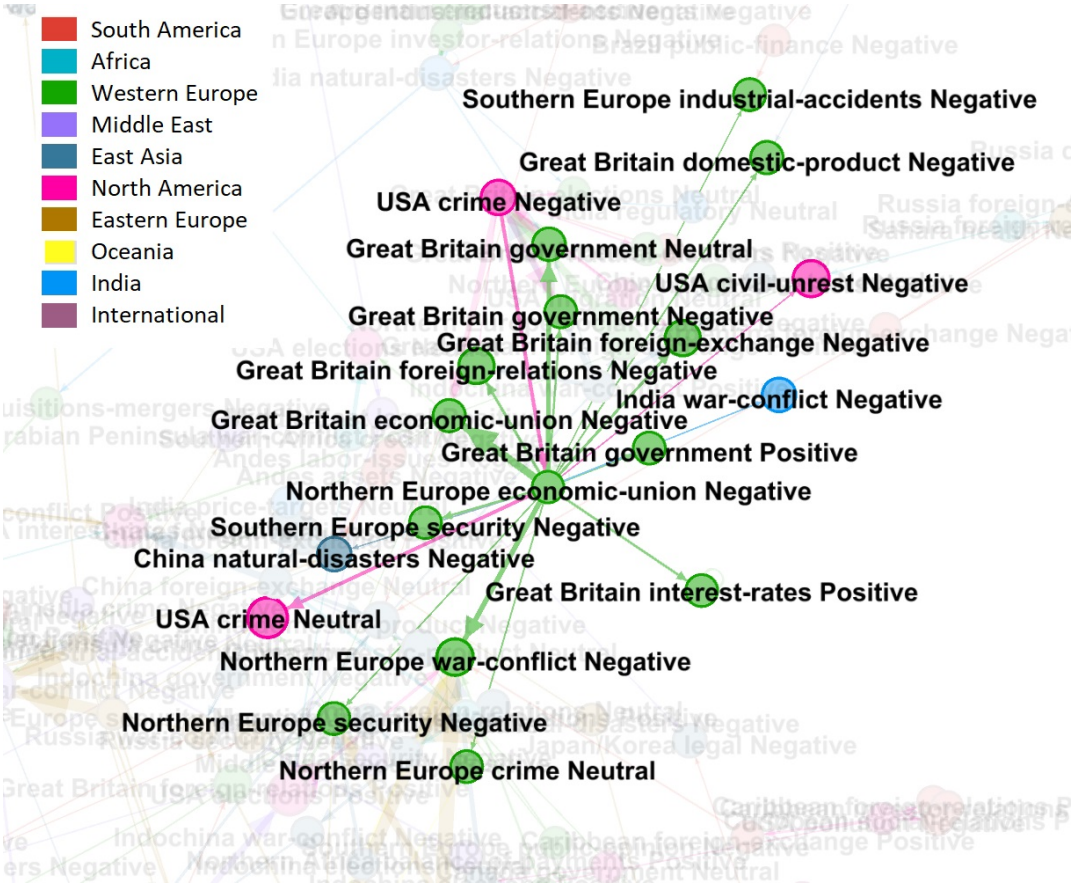
The news is a massive source of information in the world today. Every day thousands of articles describe the world we live in. This information is very hard to use because of its complexity however. This is why an automatic method to gather information from the news flow can be instrumental in many applications. In this thesis methods to estimate the amount of news that will be written is proposed.

A news article is a very complex source of information. All articles' texts are unique even though they describe similar events. Also the events themselves can sometimes be very much alike another earlier event or a unique time stamp in recent history. The company RavenPack has provided a database of articles from all around the world where the unique texts are read through by a computer and sorted into topics and regions in the world.

In this thesis a model has then been tested that will estimate the amount of news that will arrive in a certain region and topic. This is done by making the model learn by historic news feed and then we can use it on future news. The expectation of the amount of news that will be produced in a certain area can reflect the happenings in the real world and may be used to see crime waves in a region or identify a financial crisis.

The learning of the model corresponds to finding relations between the correct regions and topics in the news flow. This is the primary result in this thesis. It is possible to select a single topic in a region and see what it affects and what affects it.

Finally it can be noted that the model used in this thesis is called a Hawkes model. This model has historically been used to predict earthquakes and re-tweet waves on twitter. This is because of the model's nature which is that a single event may cause several events in the aftermath. In the case of earthquakes an earthquake is often followed by another in the same general area. If a famous twitter account is posting something, there will be a re-tweet flood afterwards. In the news flow an event like Brexit will create an increased amount of news for a time. The results can be presented with a map of nodes and arrows between where the size corresponds to the importance. In the figure below one such node map from the region Northern Europe and topic Economic Union Negative can be seen. This topic is almost only written about during Brexit the last few years and as expected mostly economic topics in Europe are affected by Brexit.



A visualisation of the relations from Northern Europe Economic Union Negative. This group has a couple of weeks during Brexit where the news flow explodes.

Acknowledgements

Writing this thesis in collaboration with Lynx Asset Management has been very inspiring. I would like to thank my supervisors and quantitative analysts at Lynx Phd. Martin Rehn, Mr. Ola Backman, Professor Tobias Rydén and Phd. Jing Fu. Their proficiency in statistical modelling and continuous advice during the project were very helpful throughout the project.

I would also like to thank Professor Magnus Wiktorsson at Lund University for providing insightful help along the way and Erik Alpsten from KTH Royal Institute of Technology with whom this project has been conducted. Finally my family and Ida Wagnström has been supporting me all the way. Thank you very much.

Disclaimer

This thesis was part of a collaborative project made by me, Carl Brishammar from Lund University, and Erik Alpsten from KTH Royal Institute of Technology. We have investigated different aspects of the news flow. The collaboration made it possible to write the two thesis in a timely manner. However during the early stages of the project most of the work was done together which can be seen in the theoretical parts of the thesis, where some sections will be very much alike.

Contents

Chapter 1

Introduction

1.1 Background

Every day we get millions of news articles, press releases and blog posts. This is a massive source of information. These articles and reports can be used to understand what is happening in the different parts of the world and possibly what is to come as well. Because of the complex structure of the news flow though it is in general hard to extract any valuable information and the amount of data available makes manual methods impractical.

If the information from the news flow could be extracted though, it is reasonable to assume that there are a lot of applications. This thesis is written in collaboration with the hedge fund Lynx Asset Management and financial applications to this information are not far fetched to think about. But understanding the news flow could prove useful in other areas as well.

When doing a statistical analysis of something that is not directly measurable, such as articles, the first step is to get a practical representation of the article. Classifying if the article is from a certain country, if it is negative or what the topic is could be more practical fields than a free text. The next step is to aggregate the result. This aggregated signal will then be the news flow that can approximate the real world events..

In this thesis we are interested in the last part where we already have access to a few values of articles and try to model the actual flow of the news using that. For example interesting information could be if there is a big flow of positive news about America's economy right now or identifying an epidemic by finding clusters of negative health related news.

1.2 Related Work

Analysis and prediction of news data has gained an increased interest in recent years. This is partially due to the attention from the financial sector, but also from news providers, social media channels and other organisations looking to optimise their user experience and marketing efforts. The spectrum of analysis is rather wide and involves many different stages, e.g. natural language processing to interpret the text, data mining to handle large sets of information as well as a range of statistical methods to model data flows. The paragraphs below are relevant to this thesis project.

News analytics as a subject in the financial sector is discussed extensively in the book *"The Handbook of News Analytics in Finance"* [Mitra and Mitra, 2011], which presents several frameworks and techniques for handling news data as well as both its potentials and risks in predicting financial assets. Similarly, the articles [Heston and Sinha, 2017, sto, 2010] both deal with prediction of stock prices using news data. The first article uses a support vector machine approach using features extracted from financial news articles and historic stock prices, whereas the second article examines the prediction accuracy of neural networks for predicting stock returns. An interesting topic here is the time horizons within which the predictions are accurate. Finally, the article *"Applications of a multivariate Hawkes process to joint modelling of sentiment and market return events"* [Yang et al., 2017] explores the use of point processes and Hawkes processes to model events in financial markets. More specifically, the study analyses how positive and negative sentiments in news events connect to positive and negative returns in the context of multivariate Hawkes processes.

Another related area is that of modelling and prediction of events on social media, e.g. how trending content goes viral and spreads on different channels as well as how it can be used to predict events outside the social media platforms. The article [Asur and Huberman, 2010] uses data from Twitter to forecast the revenues of box-office movies. In addition, [Yu and Kak, 2012] discusses several topics within the subject, such as marketing, information validity and prediction of election outcomes. Lastly, the article *"A tutorial on Hawkes Processes for Events in Social Media"* [Rizoiu et al., 2017] provides an introduction to the concept of Hawkes processes and the self-exciting properties, with a focus on social media events.

More specifically on the topic of Hawkes processes, it has been used in for example earthquake forecasting as well as modelling epidemic outbreaks in addition to the financial applications introduced above. The common characteristic in these areas is the self-exciting property. For instance, for epidemic diseases it may be reasonable to suggest that observing a case of the disease will increase

the risk (i.e. the intensity) to see more cases in that region within the near future, thus making the Hawkes process model a suitable candidate. One such study is [Schoenberg et al., 2017], which uses Hawkes process as well as another type of point process to model measles occurrences between 1906 and 1956. Another relevant article is [Bray and Paik Schoenberg, 2013], which reviews the Hawkes process among other model alternatives for earthquake forecasting. In this context, the excitation function is modelled to take the form of the so called Omori's law, which is an important concept in seismology.

1.3 Hypothesis and Purpose

In this Thesis we aim to see patterns in the news flow. The main goal was to be able to tell how plausible it is that an article of some specified parameters are to appear in the news flow in the near future.

1.4 Outline

Chapter 1: The introductory chapter have hopefully given the reader a reason to read further and insight towards where the research frontier stands today.

Chapter 2: Chapter 2 is supposed to give the reader a quick understanding of the aim of the models. How they work and what sort of characteristics they may pick up from the news flow. No details concerning what mathematical constructions that will be made will be in this chapter.

Chapter 3: Chapter 3 will handle all things related to the data set. The data used in the thesis has to be cleaned before use. This chapter describes the raw data so that the reader can understand what sort of modification had to be done to the data.

Chapter 4: In Chapter 4 all mathematical definitions and theorems will be provided. Basic discussions relating the processes and statistical tools will also be here for readers that need some recapitulation regarding those subjects. Some statistical knowledge will be expected from the reader. The models used in this thesis will get a general introduction.

Chapter 5: Succeeding the mathematical definitions the actual details of the implementation of the models will be provided in Chapter 5. What modifications that has been done to ensure good and computationally feasible results. Both the likelihood derivation and the algorithms that are used to maximise them can be found

in this chapter.

Chapter 6: Chapter 6 handles the experiment setup. What results that can be expected in the result section can be seen as well.

Chapter 7: This chapter is exclusively for the results of the thesis. Likelihoods and BIC for the different models and intensity curves will be here. The main result from this thesis, the connections between news topics and regions will take up the main part of this chapter, presented as node graphs.

Chapter 8: This chapter will have a brief discussion of how to understand the results and how the calibration step of the models could have been done better. The last section will be an outlook towards further research and how to use the results from this project.

Chapter 9: In the Conclusion chapter there will be a summary of the results and how the results relate to the subject studied in this thesis.

Chapter 2

Modelling Background

Modelling the news flow is in extension an attempt to model the events happening in the real world. In this thesis models which treat the news articles as the whole truth will be used. This is in distinction with models that use the news articles as reports of what is happening in the world which is of course a more accurate model since that is in fact what a news report is. In a model setting this can be fit into the framework as the driving process and the observable states. In our case the news feed is both the driving process and the observable states. This is as noted above a simplification since the events in the world is not, at least not completely, dependent on what is written in the news. In this model world every article will correspond to exactly one event in the setting since the news feed is the driving process. The notation *event* further down in the thesis will therefore be used for both articles and the events they describe.

A driving process with events happening at certain time stamps has to be able to create a time span with simulated events. This is done by creating a model with certain characteristics. That can be knowledge from the world or something that depends on past data. In our models we will hypothesise that the news flow can be split up in 2 parts, the news that depend on other news and news that occur independently.

2.1 Dependent and Non-dependent News

The first part is the non-dependent news articles. These include all articles that over a given time span occurs with the same probability uncorrelated with the events in the world. One can imagine that this is *almost* true for news reporting results of a sports team or standard weather reports.

On the other hand one can also imagine that there sometimes occurs something extraordinaire with the sports team or weather in question. Maybe there is a hail storm which prevents that football match to be played in the first place. Then the expected result report from the match would not happen. This part of the feed will be taken care of by the second part of the models. This will try to capture the dependent part of the news feed.

This will be done by saying that an event in a certain topic or country has an increased probability to happen if an event has happened in another dependent topic-country combination. On a very small scale we can concretize this with an example where we have a murder report. The days following it there is a higher probability of a report of a caught murderer than if there was no murder in the first place.

Of course this is very simple with hand picked examples like this. In the actual model there has to be compromises because of data and computational power available. If the model is too fine, with a lot of topics and countries, there will not be enough observations for every topic to calibrate the model. On the other hand if the model is too crude it will be hard to know what an event in the topic is about. Also too many groups of countries and topics brings problems in terms of computational power.

The dependence will further down in the thesis this will be called *excitation*. This is because in the model used in this thesis the dependence is directed. Then one topic *excites* the other. With the excitation between the topics there is also a question of what time span the excitation is apparent. In this thesis we have chosen to let the decay of the excitation be exponentially decaying over time with a limit on the half-time of 2 years.

2.2 The Two Models

In this thesis we will compare two different models. The description above concludes the similarities between the two. There are some small differences between the models even though they are in principal based on the same ideas. In the first model, *Discrete Classes*, every news article will be assigned to a topic-country combination in which it is written. These combinations can then be clustered together into bigger groups. These groups however will contain all of the subgroups' articles.

In the second model, *Overlapping Classes*, these groups can be thought of as a multivariate density of subgroups. For example two groups might be primarily Nordic countries and Russia. News written in Finland might have big similarities with the Nordic group and a minor similarity with the Russia group. In the opposite way, both news that should affect the Nordic and Russia will affect Finland. A visual-

isation of this can be seen in figure 2.1 where the green and red graph shows how much each country is related to group Nordic(red) and group Russia(green).

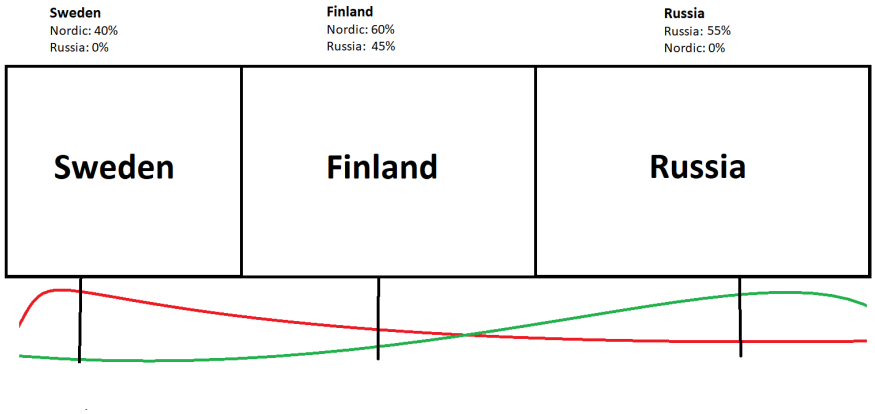


Figure 2.1: An illustration of the densities that describe the classes in the Overlapping Classes model. Two classes are here represented by the colour red and green. Their respective density can be seen in the figure. The red density contain more of the Swedish news and the green more Russian news. Both classes will contain news from Finland.

The attentive reader can in the figure see that groups contain a certain percentage of different countries/topics but every country does not necessarily have an impact in any group or impact in more than summed 100% in different groups. This will be explained in detail later when the models are fit into a mathematical frame work.

The overlapping behaviour was explained using countries since it is very intuitive but for the topics the exact same thing can be done. This gives model 2 a bit more flexibility when calibrating which hopefully results in a more precise model.

2.3 Adapting Classes

However there is an, probably more important, advantage with the second model than more flexibility which is the differentiable groups. This comes back to the intricate problem of making the groups. In the discrete model we have to customise the groups and find manually what countries that probably are affected by the same countries and affect the same countries. With differentiable groups this step could be done in the calibration of the model. In figure 2.1 this corresponds to tuning the densities.

2.4 News Flow

In this thesis there is a distinction made between the news *feed* and the news *flow*. This can be seen as the input and the output of the model. Or rather, the feed as a time series of events and the flow as the expected articles incoming in the near future, the intensity. The output news flow or intensity could then be used to simulate a fictional news feed either to reconstruct the past as good as possible or to predict the future.

However, we will never be able to reconstruct the news articles exact location in time since the intensity is a probability that an event will happen over a time span. This is why the intensity itself probably is more interesting to look at and most results will be revolving around that instead of predicting actual upcoming news.

Chapter 3

Data

This sections presents some insight into the important properties of news data in general. Before using the data in the calculations, it was important to carefully study the structure of the data as well as identify potential flaws or problems that may affect the results. Therefore, this chapter first provides some general information about the structure of news data. Thereafter, an overview of the structure and characteristics of the specific dataset used in this study is presented. Given this information, a motivation about how the data ought to be used is formulated, which in turn motivates the choices of mathematical models in the next chapter.

3.1 Dataset Overview

The dataset used throughout this study contains news data from January 1st 2000 until February 28th 2017. Even though the original source of each data point is an actual article or press release, the information available in the dataset has been processed to obtain a more compact representation. That is, each original news piece is first processed by the dataset provider and translated into their standardized framework. In this framework, each point is represented as a vector with a set of pre-specified attributes, some of which are numerical values and some categorical. One important observation here is that several data points may relate to the same original article, for instance if the article is long and contains information about different entities. The table 3.1 presents a list of some of the most important fields with corresponding descriptions.

Table 3.1: Dataset significant fields

TIMESTAMP_UTC
A date-time stamp on the form YYYY-MM-DD-hh:mm:ss.sss indicating when the news data was received by the interpreting system.
HEADLINE

The headline text of the original news article
RP_STORY_ID Unique ID to each data point in the system, distinct across all records.
ENTITY_TYPE The type of identified entity, which can be either *Commodity, Company, Currency, Nationality, Organization, People, Place, Product* or *Sports teams*.
ENTITY_NAME The name of identified entity, e.g. the name of a company or currency
ENTITY_ID Unique ID related to the identified entity, i.e. all news data points with with this entity ID is also about the same entity.
COUNTRY_CODE Two-character string with the ISO-3166 country code associated with the news data point, e.g. *US, CH, XX*.
RELEVANCE A score taking integer values between 0-100 which specifies how strongly related the identified entity is to the original article, where 0 means it was merely passively mentioned and 100 means it was considered central to the story.
EVENT_SENTIMENT_SCORE The event sentiment score states hos positive/negative that event is. More specifically, it is a score between -1 and 1 with 2 decimal places that represents the news sentiment where -1 is very negative and 1 is strongly positive.
EVENT_RELEVANCE An integer score taking values 0-100 that indicates the relevance of the identified event. A score of 100 means that it is important and stated in the headline, whereas a lower score means it was less central and stated further down in the article.
FACT_LEVEL String indicating whether the news story is considered a *fact, forecast* or *opinion*.
PROVIDER_ID The ID of the provider of the news content, e.g. *AN* for Alliance News and *DJ* for Dow Jones Newswires.
NEWS_TYPE String that specifies the type of news, e.g. *FULL-ARTICLE* or *NEWS-FLASH*.

In addition to these fields, RavenPack has a hierarchical taxonomy system to classify the content of the news data points. This particular subset of fields enables categorization and filtering on different levels of granularity. The layers in this hierarchical

structure are as presented below in Table 3.2.

Table 3.2: Dataset hierarchical fields

TOPIC Highest order in the classification. Can take either of the 5 labels: *economy, business, society, politics* or *environment*
GROUP Second level classifier, which has a total of 56 possible values, e.g. *foreign-exchange, war-conflict, acquisition-mergers, government* and *stock-prices*
TYPE A more fine-grained version of Group with 495 different labels. For instance, the group *war-conflict* has types *violence, military-action, bombing* etc.
SUB_TYPE A further subdivision of the type attribute. For instance, the group *war-conflict* has sub-type labels *attack, threat, exercise* etc.
PROPERTY An attribute of the event, such as a role or entity. For instance, the group *war-conflict* has property labels *target, attacker, location* etc.
CATEGORY The most detailed level, being a combination of sub-type and property

One key observation from the data is the existence of an event and how it affects the structure of the corresponding data point. More specifically, the data can very broadly be separated into two categories; those with an identified event and those without. The points without an event contain substantially less information and lack both sentiment score as well as the hierarchical field structure. Out of the total amount of data points, 8.4% contain an event and related information. Throughout this study, the points without an event were deemed to contain too little information and were thus left out of the analysis. That is, only points containing an event and the related information were used in order to perform the desired calculations.

3.2 Visualisations of Important Fields

In this section, a number of visualisations are presented to provide a better insight into the data characteristics and the distributions of important fields. All plots consider data with existing sentiment from all countries and throughout the entirety of the time frame.

Firstly, Figure 3.1 presents a histogram with the number of news data points from the 20 countries with the largest news flows. The countries are represented by their

two-figure country code, see appendix A for country names. One important realisation here is the large over-representation of news from USA, having a flow eight times larger than the second largest source Great Britain. In addition, the country code *XX* indicates international news that were not tied to a specific country. A total of 253 distinct country codes were present in the data set, out of which some had a very small appearance frequency.

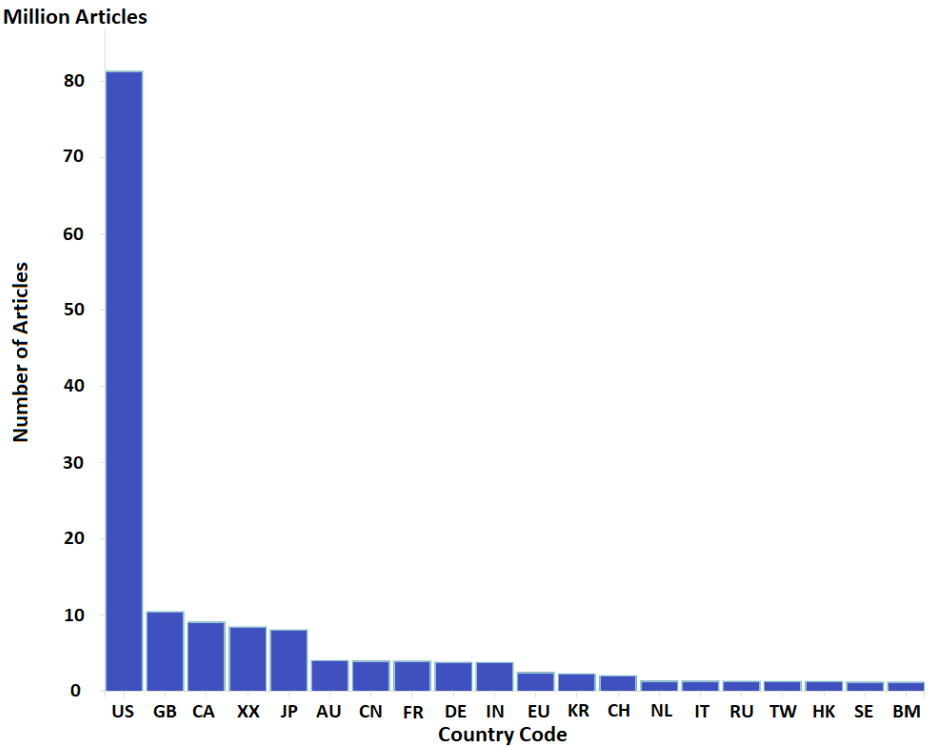


Figure 3.1: Number of articles from the 20 largest provider countries.

Next, Figures 3.2 and 3.3 shows pie charts of the distributions of relevance and event relevance fields respectively.

Distribution of the Relevance fields

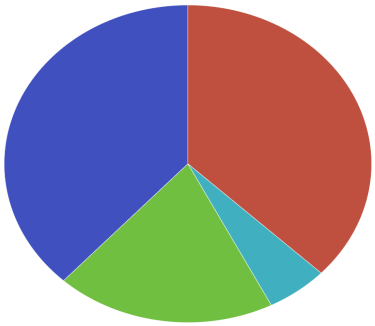


Figure 3.2: Relevance R
Red: $0 \leq R < 80$
Cyan: $80 \leq R < 90$
Green: $90 \leq R < 100$
Blue: $R = 100$

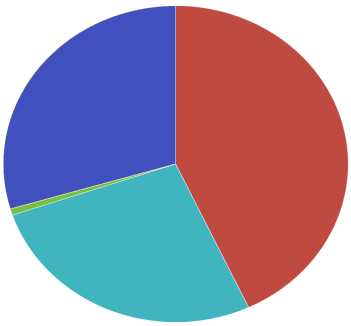


Figure 3.3: Event Relevance ER
Red: $0 \leq ER < 80$
Cyan: $80 \leq ER < 90$
Green: $90 \leq ER < 100$
Blue: $ER = 100$

Another topic worth examining is the size of the news data flow over time. Figure 3.4 presents a histogram with the yearly amount of news. Here, it is noticeable that the number of news data points has increased quite steadily over the years. One explanation for this is the increased availability of online news. Note here that the year 2017 only contains data points from the months January and February.

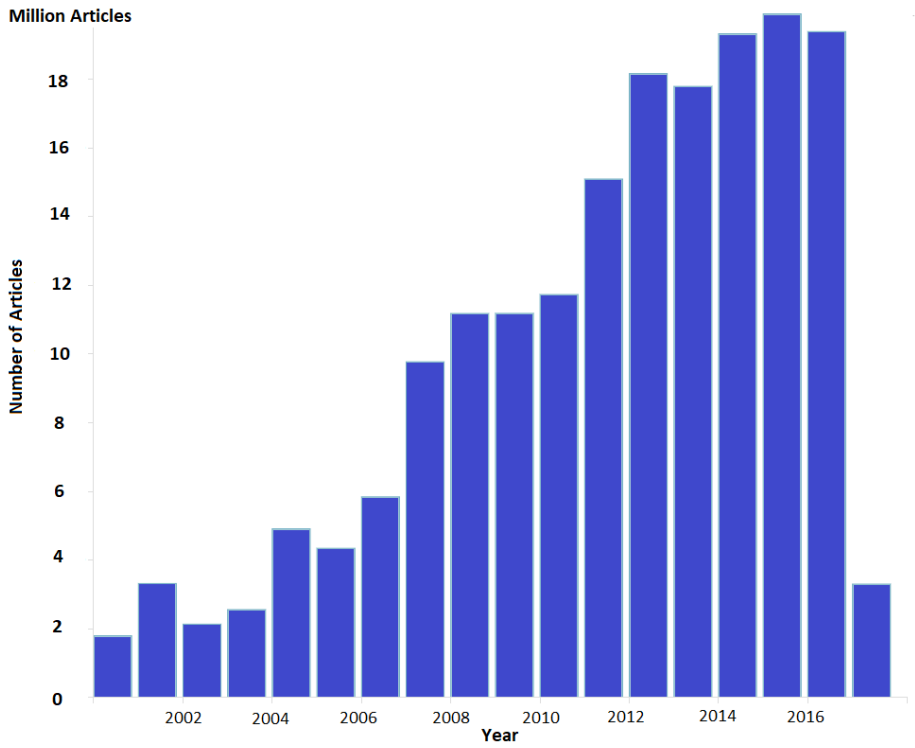


Figure 3.4: Number of articles per year Jan 2000 - Feb 2017.

Other properties can be seen when examining smaller time scopes. Figure 3.5 shows the daily amount of news for the month of December in 2014. Here, a seasonality phenomenon can be seen, where the news data flow is substantially smaller during weekends in comparison to the week days. Additionally, the flow of news is seen to decrease over the Christmas holidays, indicating that there is a seasonality component for larger holidays as well.

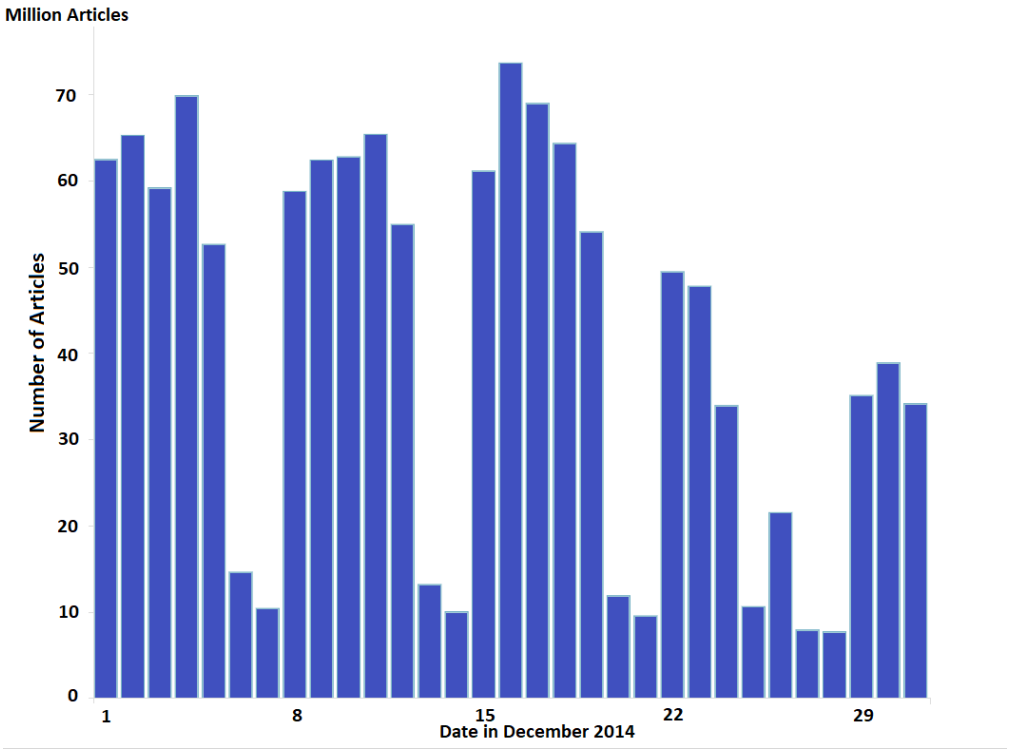


Figure 3.5: Number of articles per day in Dec 2014.

Moreover, figure 3.6 shows the news data count for a selection of groups. It can be seen that there are quite noticeable difference in the amounts, where groups like *stock-prices* have substantially higher frequency than for instance *pollution*.

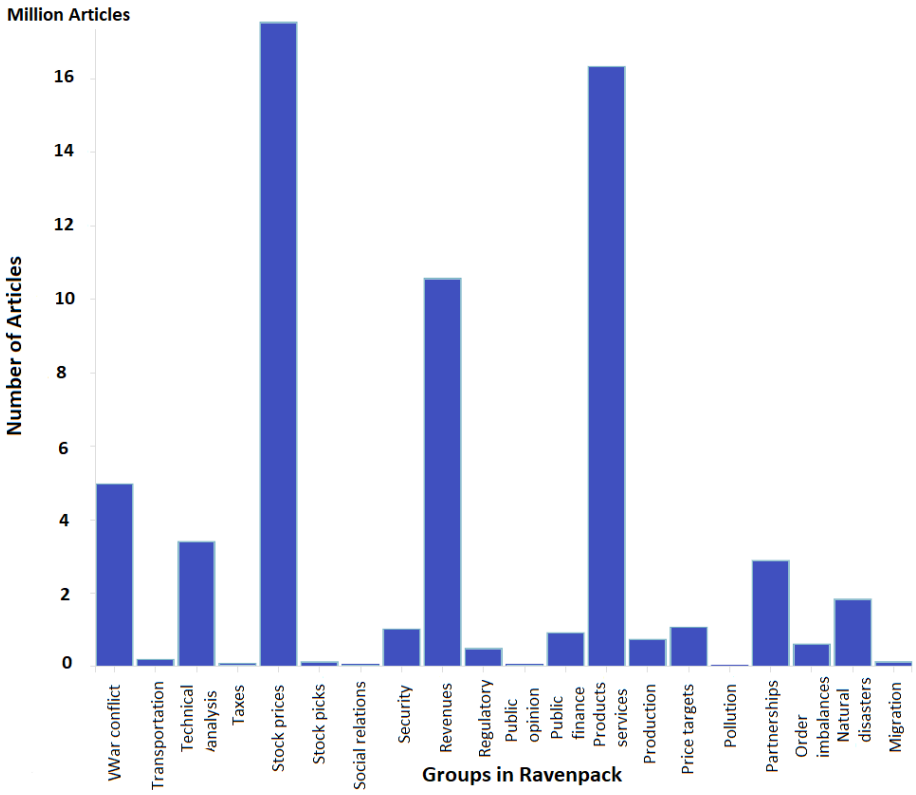


Figure 3.6: Number of articles for 20 of the 56 Group labels in the data.

Lastly, Figure 3.7 shows the distribution of the event sentiment score field with the range interval 0.05 between -1 and 1. Everything between -0.3 and 0.3 is treated as a neutral sentiment.

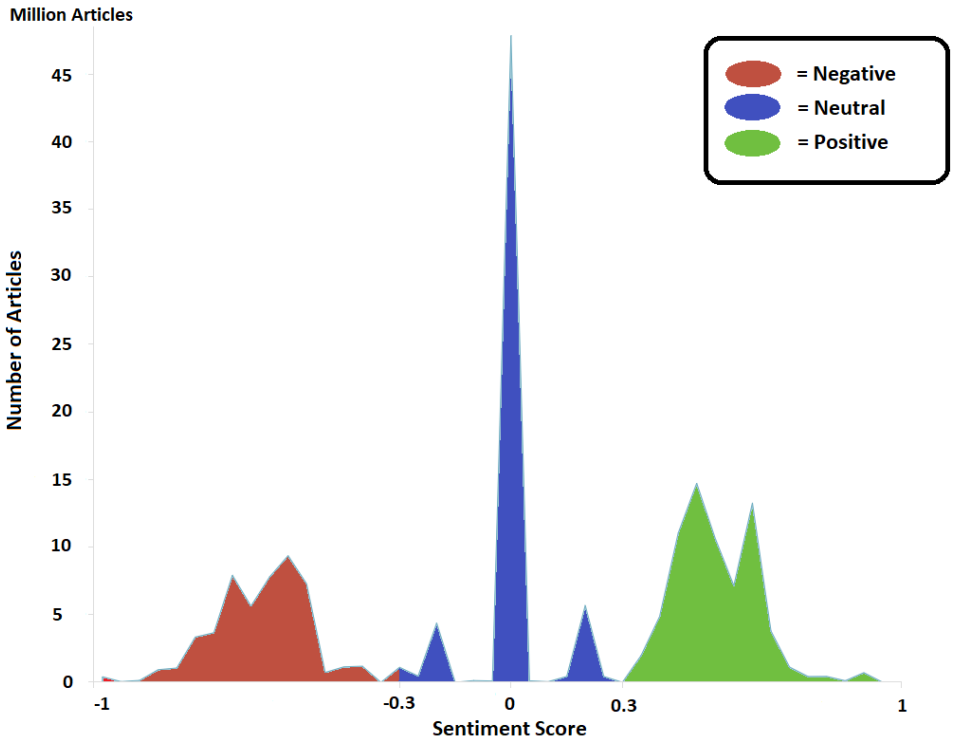


Figure 3.7: Event sentiment score distribution. A neutral article is defined as between -0.3 and 0.3 sentiment score. Lower and higher than that yields a negative and positive article respectively.

3.3 Using the Data

To get a good data set to work with out of this we want as many observations as possible split up in a convenient amount of groups. To do this we will use data from January 2007 to December 2016. This is to try to get a somewhat homogeneous flow where the internet at least existed in the beginning. Also some financial depressions and wars/conflicts has occured since then which should make the data a bit diverse.

To get as many observations as possible we will not limit anything according to relevance. That is, in figure 3.2 and figure 3.3 all observations will be used. This will of course give a higher variance to our results however when not filtering on relevance the data set goes from 30 million to 170 million observations. Our hope in this is that the increase in observations will be more significant than the increased variance because of not filtering.

For country and sentiment the respective label in Ravenpack will be used and for

topic the second layer group in the hierarchy will be used. Later in the thesis this group will be called topic occasionally since that is a more common name for it.

Chapter 4

Mathematical Background

In this chapter, a formal mathematical background of the models utilized in this thesis project is given. To begin with, some theory about stochastic processes is presented. In particular, this part focuses on presenting the Hawkes process, its important properties as well as how it is different from more simple models. Thereafter, the models for the news data flows are formalized, which ties back to both the data structure presented in Chapter 3 as well as the stochastic process theory provided in the first section of this chapter. After this, the chapter provides some background theory on the optimization algorithms and parameter estimation procedures used in the implementation. (Lastly, some background for the clustering methods used throughout the study are presented.)

4.1 Stochastic Processes

This part provides some important mathematical background on stochastic processes and the Hawkes process in particular, which is an essential part of modeling the news data flow in this thesis project. However, prior to defining the Hawkes process model, some more basic concepts are outlined.

4.1.1 Basic Stochastic Processes

Firstly, the topic of point processes is an important concept in probability theory and is especially central in modeling spatial data. In the setting of news data flow, a point process can intuitively be thought of as the random variables describing the news arrival times. The formal definition of a point process [Laub et al., 2015, Karr, 1986] is stated below.

Definition 4.1. *(Point process)*
A sequence of non-negative random variables $\mathbf{T} = \{T_1, T_2, \dots\}$ *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a point process if*

(i) $\mathbb{P}(0 \leq T_1 \leq T_2 \leq \dots) = 1$,

(ii) *The number of points in a bounded region of* $[0, \infty)$ *is finite almost surely, i.e.* $\mathbb{P}\left(\lim_{n\to\infty} T_n < \infty\right) = 1$.

In many cases, a point process has a corresponding count process that describes the cumulative count of arrivals. The definition of the counting process is presented below.

Definition 4.2. *(Counting Process)*
A stochastic process $N\colon [0,\infty) \times \Omega \to \mathbb{N}_0$ *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, with* $N_t\colon \Omega \to \mathbb{N}_0$ *such that* $N_t(\omega) = N(t, \omega)\ \forall\ \omega \in \Omega$*, is a counting process if*

(i) $\mathbb{P}(N_0 = 0) = 1$,

(ii) $\mathbb{P}(N_t < \infty) = 1,\ \forall\ t \in [0, \infty)$,

(iii) *it holds for all* $s, t \in [0, \infty)$ *with* $s < t$ *that* $\mathbb{P}(N_s \leq N_t) = 1$,

(iv) *it holds that* N *is a non-decreasing right-continuous step function with increment size 1.*

Furthermore, a useful concept related to the point- and counting processes is the history sigma algebra. That is, for each time $t \in [0, \infty)$, the history sigma algebra $\mathcal{H}_t$ of a counting process N is given as $\mathcal{H}_t = \sigma\left(\{N_u\colon 0 \leq u \leq t\}\right)$. Consequently, the sequence $\mathcal{H} = \{\mathcal{H}_t\}_{t\in[0,\infty)}$ is a filtration on the measurable space $(\Omega, \mathcal{F})$. How a counting process depends on its related filtration is of great significance in many applications and its importance for this study will be presented later. An important counting process with some special such properties is the Poisson process, which is defined below.

Definition 4.3. *(Homogeneous Poisson Process)*
A counting process $N\colon [0,\infty) \times \Omega \to \mathbb{N}_0$ *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a homogeneous Poisson process with intensity* $\lambda \geq 0$ *if for arbitrary* $t \in [0, \infty)$ *it holds for all* $h \geq 0$ *that*

$$\mathbb{P}\left(N_{t+h} - N_t = m\right) = \begin{cases} 1 - \lambda h + \mathcal{O}(h), & m = 0, \\ \lambda h + \mathcal{O}(h), & m = 1, \\ \mathcal{O}(h), & m > 1, \end{cases} \tag{4.1}$$

where $\mathcal{O}$ signifies some function $o\colon [0,\infty) \to \mathbb{R}$ with the property

$$\lim_{h \searrow 0} \frac{o(h)}{h} = 0, \tag{4.2}$$

which also gives that $o(0) = 0$. This definition in turn implies that non-overlapping intervals of N are independent random variables, i.e. for all combinations $s, t \in [0, \infty)$ such that $s < t$ it holds that the increment $N_t - N_s$ is independent of $\mathcal{H}_s$. Moreover, the increments are stationary and $N_t - N_s \sim \text{Po}(\lambda(t-s))$. The term *homogeneous* specifies that there is no time dependency in the intensity, However in some situations it may happen that the intensity is not a constant but instead vary with time, e.g. with some linear increase or seasonal oscillations. In such a case, an *inhomogeneous* Poisson process is obtained. The definition of such a process is presented below.

Definition 4.4. *(Inhomogeneous Poisson process)*
A counting process $N\colon [0,\infty)\times\Omega \to \mathbb{N}_0$ *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is an inhomogeneous Poisson process with intensity function* $\lambda\colon [0,\infty) \to [0,\infty)$ *if for arbitrary* $t \in [0,\infty)$ *it holds for all* $h \geq 0$ *that*

$$\mathbb{P}\left(N_{t+h} - N_t = m\right) = \begin{cases} 1 - \lambda(t)h + \mathcal{O}(h), & m = 0, \\ \lambda(t)h + \mathcal{O}(h), & m = 1, \\ \mathcal{O}(h), & m > 1, \end{cases} \tag{4.3}$$

where as in the homogeneous case, $\mathcal{O}$ signifies some function $o\colon [0,\infty) \to \mathbb{R}$ satisfying the property in equation 4.2. For this case, it is given that

$$N_{t+h} - N_t \sim \text{Po}\left(\int_t^{t+h} \lambda(u)\,\mathrm{d}u\right), \quad t \in [0,\infty). \tag{4.4}$$

4.1.2 The Hawkes Process

Now, it is time to formally introduce the Hawkes process, which is a fundamental part in this thesis study. The Hawkes process is in some ways a generalization of the Poisson process, however where the process is self-exciting. This means that every observed arrival in the process causes an increase in the value of the intensity function, thus also increasing the probability of observing more arrivals in the future. In addition, this implies that the intensity does not only vary with time, but also depends on the history sigma algebra generated by the process up until the current time point. The definition of the Hawkes process is thus presented below.

Definition 4.5. *(Hawkes process)*
A counting process $N\colon [0,\infty) \times \Omega \to \mathbb{N}_0$ *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *with associated filtration* $\mathcal{H}$ *is a Hawkes process if for arbitrary* $t \in [0,\infty)$ *it holds that*

(i) for all $h \geq 0$

$$\mathbb{P}\left(N_{t+h} - N_t = m \,|\, \mathcal{H}_t\right) = \begin{cases} 1 - \lambda^*(t)h + \mathcal{O}(h), & m = 0, \\ \lambda^*(t)h + \mathcal{O}(h), & m = 1, \\ \mathcal{O}(h), & m > 1, \end{cases} \tag{4.5}$$

(ii) the conditional intensity function λ^* *is given as*

$$\lambda^*(t) = b + \int_0^t \nu(t-u)\,\mathrm{d}N_u, \quad t \in [0, \infty) \tag{4.6}$$

where $b > 0$ *is defined as the background intensity and* $\nu \colon (0, \infty) \to [0, \infty)$ *is defined as the excitation function.*

As before, $\mathcal{O}$ signifies some function $o \colon [0, \infty) \to \mathbb{R}$ satisfying the property in equation 4.2. Here, the conditional intensity function is an important difference from the previous Poisson process since it depends on the history of the process and so its future values are not deterministic. More formally, the conditional intensity function λ^* can be defined as

$$\lambda^*(t) = \lim_{h \searrow 0} \frac{\mathbb{E}[N_{t+h} - N_t \,|\, \mathcal{H}_t]}{h}, \quad t \in [0, \infty). \tag{4.7}$$

Furthermore, the choice of excitation function ν may vary between applications and used data. One popular choice that has been used in for example seismological modelling is a function, also called Omori's law, on the form

$$\nu(t) = \frac{k}{(c+t)^p}, , \quad t \in [0, \infty), \tag{4.8}$$

where k, c, p are positive constants. Another common option is an exponential kernel on the form

$$\nu(t) = Ve^{-\gamma t}, , \quad t \in [0, \infty), \tag{4.9}$$

where V, γ are some positive constants. Moreover, it can be noted that if $\nu(t) = 0 \;\forall\; t \in (0, \infty)$, the Hawkes process becomes identical to the homogeneous Poisson process. For an observed sequence of arrival times $\mathbf{t} = \{t_1, t_2, \dots\}$ of the process during a time interval $[t_a, t_b] \subset [0, \infty)$, the conditional intensity function presented in equation 4.6 can be written as

$$\lambda^*(t) = b + \sum_{\substack{t_l \in \mathbf{t}: \\ t_l < t}} \nu(t - t_l), \quad t \in [0, \infty). \tag{4.10}$$

Consequently, the likelihood function and corresponding log-likelihood function of such a realisation can be written as

$$\mathcal{L} = \prod_{t_l \in \mathbf{t}} \lambda^*(t_l) e^{-\int_{t_a}^{t_b} \lambda^*(u)\,\mathrm{d}u}, \tag{4.11}$$

$$\log \mathcal{L} = \sum_{t_l \in \mathbf{t}} \log\left(\lambda^*(t_l)\right) - \int_{t_a}^{t_b} \lambda^*(u)\,\mathrm{d}u. \tag{4.12}$$

The proof for deriving this likelihood is left out of this report, however a derivation of the expression can be found in the literature reference [Laub et al., 2015]. Next, the first Hawkes process can be extended to the case where multiple counting processes are considered. In such a case, the processes can have both self- and mutually-exciting properties, i.e. each process' intensity is not only influenced by itself but also by the other counting processes. Such a scenario can be modeled using the multivariate Hawkes process, which is defined below.

Definition 4.6. *(Multivariate Hawkes Process)*
Consider a collection of n counting processes $\mathbf{N} = \{N^{(1)}, \ldots, N^{(n)}\}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with associated filtration $\mathcal{H}$. Then $\mathbf{N}$ is a multivariate Hawkes process if for each $i \in \{1, \ldots, n\}$ it holds that

(i) for all $h \geq 0$

$$\mathbb{P}\left(N_{t+h}^{(i)} - N_t^{(i)} = m \,|\, \mathcal{H}_t\right) = \begin{cases} 1 - \lambda_i^*(t)h + \mathcal{O}(h), & m = 0, \\ \lambda_i^*(t)h + \mathcal{O}(h), & m = 1, \\ \mathcal{O}(h), & m > 1, \end{cases} \tag{4.13}$$

(ii) the conditional intensity function λ_i^* corresponding to N^i can be written on the form

$$\lambda_i^*(t) = b_i + \sum_{j=1}^{n} \left(\int_0^t \nu_{ij}(t - u)\,\mathrm{d}N_u^{(j)} \right), \quad t \in [0, \infty), \tag{4.14}$$

where b_i is the background intensity function and $\nu_{ij} \colon (0, \infty) \to [0, \infty)$ is the excitation function from $N^{(j)}$ to $N^{(i)}$.

As before, $\mathcal{O}$ signifies some function $o\colon [0,\infty) \to \mathbb{R}$ satisfying the property in equation 4.2. Next, consider an observed sequence of arrival times $\mathbf{t}^i = \{t_1^i, t_2^i, \dots\}$ corresponding to each counting process $N^{(i)}, i \in \{1, \dots, n\}$ during a time interval $[t_a, t_b] \subset [0,\infty)$. The conditional intensity function λ_i^* for each i can thus be written as

$$\lambda_i^*(t) = b_i + \sum_{j=1}^{n} \sum_{\substack{t_l^j \in \mathbf{t}^j \colon \\ t_l^j < t}} \nu_{ij}(t - t_l^j), \quad t \in [0,\infty), \tag{4.15}$$

with the likelihood function and corresponding log-likelihood taking the forms

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{t_l^i \in \mathbf{t}^i} \lambda_i^*(t_l^i) e^{-\int_{t_a}^{t_b} \lambda_i^*(u)\,\mathrm{d}u}, \tag{4.16}$$

$$\log \mathcal{L} = \sum_{i=1}^{n} \left(\sum_{t_l^i \in \mathbf{t}^i} \log\left(\lambda_i^*(t_l^i)\right) - \int_{t_a}^{t_b} \lambda_i^*(u)\,\mathrm{d}u \right), \tag{4.17}$$

i.e. the total likelihood is a product over terms similar to those presented in equation 4.11. Additionally, the exponential excitation function introduced in equation 4.9 can be extended to the multivariate case to model the excitation from $N^{(j)}$ to $N^{(i)}$ using the form

$$\nu_{ij}(t) = V_{ij} e^{-\gamma_j t}, \tag{4.18}$$

where V_{ij}, γ_j are constants, which inserted in equation 4.14 gives the conditional intensity function for each i to take the form

$$\lambda_i^*(t) = b_i + \sum_{j=1}^{n} \sum_{\substack{t_l^j \in \mathbf{t}^j \colon \\ t_l^j < t}} V_{ij} e^{-\gamma_j (t - t_l^j)}, \quad t \in [0,\infty) \tag{4.19}$$

Here, V_{ij} can be thought of as elements in an excitation amplitude matrix $V \in \mathbb{R}^{n \times n}$. Similarly, the parameters b_i and γ_j and can be thought of as elements in vectors $b \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}^n$ respectively. Of course, alternative expressions for the excitation function can also be proposed, e.g. by stating a non-stationary model where the parameters can vary with time. For instance, by redefining the background intensity constant b_i as a function $b_i\colon [0,\infty) \to [0,\infty)$ a case similar to the one with the inhomogeneous Poisson processes presented in Definition 4.4 is obtained. This will further down be called an Inhomogeneous Hawkes Process.

4.2 Modeling News Data

This section provides a description for how the news data is modeled throughout this study. To begin with, every news data point observed during some time interval $[t_a, t_b] \subset [0, \infty)$ is represented by a point y_i such that

$$y_i = (t_i, x_i) \in [t_a, t_b] \times \mathcal{X}, \tag{4.20}$$

where t_i is time stamp at which the piece of news was observed, x_i is the set of attributes assigned to the point by the text interpretation system and $\mathcal{X}$ is the attribute space. For instance, if the data point is described with m real-valued numerical attributes it is obtained that

$$x_i \in \mathcal{X} \subseteq \mathbb{R}^m. \tag{4.21}$$

Next, the sequence of observed data is defined by $\mathbf{y} = \{y_1, y_2, \dots\}$ with associated arrival times and attributes defined as $\mathbf{t} = \{t_1, t_2, \dots\}$ and $\mathbf{x} = \{x_1, x_2, \dots\}$ respectively. This data sequence includes the whole set of observed news data points, i.e. there is no sorting process based on the content of news. However, in order to properly apply the multivariate Hawkes process model, the aggregated news data ought to be partitioned into classes where each class is characterized by containing homogeneous types of news. In this study, the model for partitioning the news flow into classes have two different versions; *distinct classes* and *overlapping classes*. In short, distinct classes here means that the attribute space $\mathcal{X}$ is divided into distinct subsets, where each subset corresponds to a class, whereas overlapping classes means that each class is represented by a probability density over the whole attribute space. Illustrations of these two concepts are presented below in Figures 4.1 and 4.2. Both of these two model alternatives are also outlined more in detail in the next subsections.

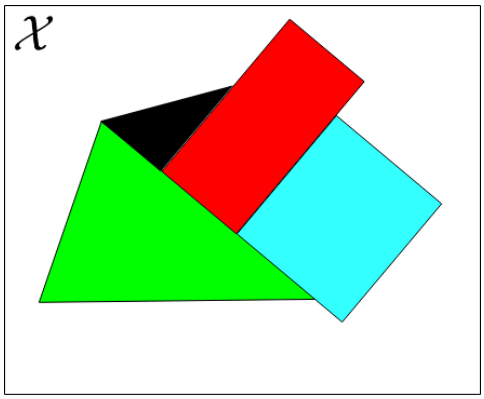


Figure 4.1: Distinct classes

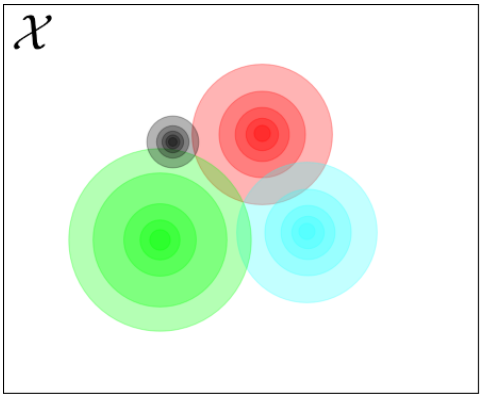


Figure 4.2: Overlapping classes

4.2.1 Distinct Classes

In the first model it is assumed that the attribute space $\mathcal{X}$ is separated into disjoint classes. That is, if there is a total of n classes it is assumed that

$$\mathcal{X} = \bigcup_{i=1}^{n} \mathcal{X}_i, \quad \mathcal{X}_i \cap \mathcal{X}_j = \emptyset, i \neq j \tag{4.22}$$

where $\mathcal{X}_i$ is the part of the attribute space corresponding to class i. With this assumption, the flow of news data from each class i is denoted as the sequence $\mathbf{y}^i = \{y_1^i, y_2^i, \dots\}$ with associated time sequence $\mathbf{t}^i = \{t_1^i, t_2^i, \dots\}$ and attribute sequence $\mathbf{x}^i = \{x_1^i, x_2^i, \dots\}$, similarly as in the general model but here separated by class, i.e. $\mathbf{y}^i = \{y_j \in \mathbf{y} \colon\ x_j \in \mathcal{X}_i\}$.

The news flow is modeled with a multivariate Hawkes process $\mathbf{N} = \{N_1, \dots, N_n\}$, as presented in the previous section, such that each class i is represented by a counting process N_i modeling the arrival times and cumulative count of news data points in that specific class. Furthermore, by using the likelihood expression stated in equation 4.16 as well as the generalized exponential excitation function for multivariate Hawkes processes introduced in equation 4.18 with the parameters b, V, γ in the conditional intensity functions, the likelihood for the arrival times $\mathbf{t}$ of an observed news data sequence $\mathbf{y}$ during the time interval $[t_a, t_b]$ is given by

$$p(\mathbf{t}|b, V, \gamma) = \prod_{i=1}^{n} \prod_{t_l^i \in \mathbf{t}^i} \lambda_i^*(t_l^i|b, V, \gamma) e^{\left(-\int_{t_a}^{t_b} \lambda_i^*(t|b,V,\gamma)\,\mathrm{d}t\right)}, \tag{4.23}$$

where $\lambda_i^*(t|b, V, \gamma)$ indicates the function in equation 4.19 with the specific parameter choice b, V, γ.

Next, the spatial attributes of a news data points generated in class i is determined by a probability density function $f_i : \mathcal{X}_i \to [0, \infty)$. Since they are categorical and probability attributes, this density can be partitioned into a product of multinomial densities. For each class i, the multinomial parameters can be denoted by ρ_i, which contain the point-wise probabilities in the categorical domain.

Using this density model for the spatial attributes, the likelihood for an attribute sequence $\mathbf{x}$ corresponding to an observed news data sequence $\mathbf{y}$ can be written as

$$p(\mathbf{x}|\rho) = \prod_{i=1}^{n} \prod_{x_l^i \in \mathbf{x}^i} f_i(x_l^i|\rho_i), \tag{4.24}$$

where $\rho = \{\rho_i\}_{i=1}^n$. Next, it is modeled that the prior distribution for the parameters b, V, γ, ρ can be factorized such that

$$\begin{aligned} f_{bV\gamma\rho}(b, V, \gamma, \rho) &= f_{bV\gamma}(b, V, \gamma) f_\rho(\rho) \\ &= f_V(V) \prod_{i=1}^{n} f_{b_i}(b_i) f_{\gamma_i}(\gamma_i) f_{\rho_i}(\rho_i). \end{aligned} \tag{4.25}$$

Finally, given the parameterization of the time- and space factors for the news data flow as well as the factorization of the parameters' prior distribution, the likelihood function for the observed news data sequence $\mathbf{y}$ also factorizes and can be written as

$$p(\mathbf{y}|b, V, \gamma, \rho) = p(\mathbf{t}|b, V, \gamma) p(\mathbf{x}|\rho). \tag{4.26}$$

This gives the posterior distribution over the parameters to be

$$\begin{aligned} p(b, V, \gamma, \rho|\mathbf{y}) &= \frac{p(\mathbf{y}|b, V, \gamma, \rho) f_{bV\gamma\rho}(b, V, \gamma, \rho)}{p(\mathbf{y})} \\ &= \frac{p(\mathbf{t}|b, V, \gamma) f_{bV\gamma}(b, V, \gamma)}{p(\mathbf{t})} \frac{p(\mathbf{x}|\rho) f_\rho(\rho)}{p(\mathbf{x})}. \end{aligned} \tag{4.27}$$

From this expression, it can be concluded that the distinct class model with the presented properties yields the time- and attribute aspects to be separated in the posterior distribution, which in turn means that the time parameters and attribute parameters can be optimised independent of each other.

4.2.2 Overlapping Classes

A generalization of the first model would as mentioned be to no longer require the attribute space to be separated into disjoint classes. In such a case, the conditional intensity function related to the Hawkes process is redefined as a function $\lambda^* \colon [0, \infty) \times \mathcal{X} \to [0, \infty)$ such that for a sequence of data $\mathbf{y} = \{y_1, y_2, \dots\}$ observed during the time interval $[t_a, t_b]$ it is given that

$$\lambda^*(t, x) = b(x) + \sum_{\substack{(t_l, x_l): \\ t_l < t}} \nu(t - t_l, x, x_l), \quad t \in [t_a, t_b], \ x \in \mathcal{X}. \tag{4.28}$$

In this setting, how much the news flow at a point $x \in \mathcal{X}$ is influenced by other observations is determined by functions $g^i \colon \mathcal{X} \to [0, \infty)$, $i \in \{1, \dots, n\}$, such that each g^i is a density function that represents a class in this new setting, which ties back to the structure introduces in Figure 4.2. Taking the sum over these densities, a function $g \colon \mathcal{X} \to [0, \infty)$ is defined such that

$$g(x) = \sum_{i=1}^{n} g^i(x), \quad x \in \mathcal{X}. \tag{4.29}$$

Furthermore, it is modeled that the parts in the conditional intensity function take the forms

$$b(x) = \sum_{i=1}^{n} b_i g^i(x), \quad x \in \mathcal{X}, \tag{4.30}$$

and

$$\nu(t - t', x, x') = \sum_{i=1}^{n} \sum_{j=1}^{n} V_{ij} g^i(x) \frac{g^j(x')}{g(x')} e^{-\gamma_j (t - t')}, \tag{4.31}$$

where $b_i, V_{ij}, \gamma_i \in \mathbb{R}$ are constants. Each element in the first sum can be interpreted as how a point x in class i is affected by an observed data point x', weighted by the probability $\frac{g^j(x')}{g(x')}$ that point a x' is in class j.

As shown in Figure 4.2, a point in the attribute space can be contained in several classes with different probabilities. This probability will in a general context with both real and categorical variables such that for each i it holds that

$$g^i(x) = f(x|\rho_i), \quad x \in \mathcal{X}. \tag{4.32}$$

where $\rho = \{\rho_i\}_{i=1}^n$ represent the multinomial parameters describing the distribution over the categorical variables. It can be noted that this setup is the same as in the distinct classes model. However, even though the overlapping and distinct models show similarities when written this way, there are important differences. A significant difference is that the spatial and time dependent parts of the likelihood no longer are independent. The likelihood $p(\mathbf{y}|b, V, \gamma, \rho)$ for an observed news data sequence $\mathbf{y}$ has to be written in the full form as

$$\prod_{y_l \in \mathbf{y}} \lambda(t_l, x_l|b, V, \gamma, \rho) e^{-\int_{t_a}^{t_b} \int_{x \in \mathcal{X}} \lambda(t, x|b, V, \gamma, \rho) \, \mathrm{d}x \, \mathrm{d}t}, \tag{4.33}$$

and the posterior distribution of the parameters is obtained to be

$$p(b, V, \gamma, \rho|\mathbf{y}) = \frac{p(\mathbf{y}|b, V, \gamma, \rho) f_{bV\gamma\rho}(b, V, \gamma, \rho)}{p(\mathbf{y})}. \tag{4.34}$$

Now as an appetiser for the next section it can be noted that the weighting of the model is what makes this model more flexible. The gradient descent methods that are used to estimate the excitation and background intensities can in this framework be used on the multinomial distributions. That is not possible in the distinct class version where the class optimisation would be a combinatorical problem which is time consuming to the point where it is impossible.

4.3 Optimization and Parameter Estimation

To estimate the parameters in the mathematical expressions that model the news data flow was central in this thesis. More specifically, given the observations in the provided dataset and the underlying model, the likelihood function for the observed sequences could be formulated. Having stated this function, the parameters could be estimated by maximizing the likelihood with respect the these parameters in a maximum-a-posteriori manner. However with the complex models used throughout this study, closed-form solutions for the parameters could not be formulated. In addition, a large number of parameters needed to be estimated simultaneously and the size of the input data was generally very large. Dealing with big datasets as well as high-dimensional parameter spaces was therefore of vital importance. Hence, iterative methods had to be used to numerically optimize the likelihood and estimate the desired parameters. This section provides some information about these numerical methods used to estimate the parameters.

4.3.1 Gradient Descent

The gradient descent method is one of the most basic methods in numerical optimization. This subsection provides a description of the algorithm in a general setting. Consider the problem of minimizing an objective function $F\colon \mathbb{R}^m \to \mathbb{R},\ m \in \mathbb{N}$. That is, the goal is to identify an optimal solution $w^* \in \mathbb{R}^m$ such that $F(w^*) \leq F(w), \forall\, w \in \mathbb{R}^m$. Note that this is analogous to maximizing $-F$. In general, a closed-form solution for w^* can not be derived. In such a case, the gradient descent algorithm can be used to find an estimate for w^*. This algorithm requires F to be differentiable and can be described by the following steps

a. Define a prior guess w_0 for the minimum of the objective function.

b. Given current iteration index t, update the estimated solution by

$$w_{t+1} = w_t - \eta_t \nabla F(w_t) \tag{4.35}$$

where η_t is the learning rate in iteration t.

c. Check convergence criteria $|F(w_{t+1}) - F(w_t)| < \delta$ for some predefined small constant δ. If convergence criteria satisfied, break the algorithm. Otherwise, repeat from point 2.

Here, the learning rate η_t can be defined as a function of t. This rate is of importance and has to be tuned to the specific problem in question in order to produce a solution that converges to the optimal value. For a suitable choice of the learning rate, the solution is guaranteed to converge to a local minimum. However, the objective function F is in general not convex, which means that the obtained approximate local minimum is not necessarily the global minimum. Consequently, the obtained solution will often heavily depend on the prior guess w_0.

In some problems it is of usage to adapt the algorithm to the problem-specific geometry in order to achieve faster convergence. An example of such an algorithm is ADAM [Kingma and Ba, 2014], which can be seen as an extension to gradient descent that uses a cumulative gradient as well as an estimate for the second moment. This method will be used exclusively in the thesis.

4.4 Statistical Model Evaluation

When evaluating a statistical model there are some things to consider. This includes both assessment for how well a specific model fits provided data as well as comparison of hierarchical models. Typically, this becomes a trade-off between choosing a more complex model, which can adapt to the data more flexibly but may cause computational issues and overfitting, or choosing a simpler model, which may be more easily handled but provide a worse fit. This is an important part of this study. Hence, this section provides some mathematical background to the statistical evaluation tests that were utilised to compare the mathematical models examined in this study.

Firstly, perhaps the most fundamental concept in this area is the likelihood function, which has been used earlier in this report, e.g. in equations 4.11 and 4.16. Here, given the suggested underlying model and a set of observations, the likelihood function can be stipulated. More formally, the likelihood function can be presented in the following way:

Consider a collection of parameters θ for a suggested underlying model $\mathcal{M}(\theta)$. Let the random variables $X_1, \ldots, X_k$ have a joint density function $f(X_1, \ldots, X_k|\mathcal{M}(\theta))$ based on this model. For a given sequence of observations $X_1 = x_1, \ldots, X_k = x_k$ the likelihood function $\mathcal{L}$ is given as

$$\mathcal{L} = f(x_1, \ldots, x_k|\mathcal{M}(\theta)). \tag{4.36}$$

This means that $\mathcal{L}$ is the likelihood of the observations given that the model is true. Using this formulation, $\mathcal{L}$ can be thought of as a function of θ and thus maximised with respect to these parameters. Note that maximising the likelihood function $\mathcal{L}$ is

analogous to maximising $\log \mathcal{L}$ of the monotonous logarithmic function, or likewise to minimise $-\log \mathcal{L}$. This is also the procedure used to estimate the model parameters in this study, as presented in Sections 4.1-4.3 above. However, the maximised likelihood function $\widehat{\mathcal{L}}$ is not necessarily the best measure of assessment and can not always be used to compare different models. This is the case since a larger model with additional parameters has more flexibility and will therefore always yield a larger likelihood for the same set of data. For such a case, the likelihood function says nothing about overfitting. Hence, regularisation terms can be introduced to take this into account when evaluating the statistical models.

Two such alternatives which include regularisation terms are the AIC and BIC measures [Akaike, 1973, Schwarz, 1978]. Consider a dataset containing k observations and a model with q parameters. Given the maximised likelihood function $\widehat{\mathcal{L}}$ obtained from the optimisation step, the AIC and BIC measures are defined as

$$AIC = -2 \log \widehat{\mathcal{L}} + 2q, \tag{4.37}$$

$$BIC = -2 \log \widehat{\mathcal{L}} + \log(k)q. \tag{4.38}$$

Here, the BIC measures takes into account the number of observation in the input dataset and thus more heavily penalizes model complexity in the cases when $k > \exp(2)$. Thus, in the context of this study, the negative log-likelihood and BIC measures will be the primary statistics in assessment and selection of models.

Chapter 5

News Flow Models

With the theory down we are ready to finalise how the models will come together with the data seen in chapter 3.

5.1 Bucketing

To be able to actually go through the massive amount of data, the estimation of the parameters has to be done very efficiently. The data comes at a resolution of 0.001 seconds which means that there are over 10^{10} time stamps for the whole data set over 10 years. Over this time there are also about 170 million actual events recorded that are classified with a sentiment by RavenPack.

The structure of Hawkes processes are heavily dependant on old values. Becuase of this the implementation in this thesis uses recursive computation, see further down. However even though this recursive trick is used, the nature of a recursive model makes looping over old values unavoidable. Therefore every iteration of the loop has to be done in sequence. This makes parallelism of the whole program impossible and therefore that time consumption runs away quickly.

Our solution to this was to zoom out and lower the resolution of the data. To get reasonable run times the final resolution used was 1 week. This means that every article produced in the same week got the same time stamp. This is a reduction of time consumption approximately with a factor $\frac{170000000}{52*10} \approx 330000$. This comes from the amount of events divided by the time steps of the new resolution. The reason why it is possible to get this fast computation with lower resolution is because of bucketing. If the amount of events are greater than the number of time stamps it is faster to use a representation of the data as "the number of articles in a day" than the time stamp of every article. With a few thousand optimisation steps

needed to get optimised values this takes a few hours or days depending on model. Using a higher resolution would have been possible if the data set was cut to a way shorter time period or cutting the data in other ways, for example only looking at a single news type.

However, having a higher resolution is not something that is necessarily good for the interpretation of the results. In our models the excitation kernel only consists of 1 decay time per class. This has some implications. Compare that to another example model

$$v^{ij}(t) = (e^{-\gamma^{1j}t} + e^{-\gamma^{2j}t} + e^{-\gamma^{3j}t} + e^{-\gamma^{4j}t})V^{ij} \tag{5.1}$$

where it is possible to find the 4 most important decay times. With the model used in this thesis the only decay time found will be the most important one. That is, we will neglect everything that are not the most important excitation.

What happens when bucketing the articles in the same bucket is that excitation between these articles are neglected. This means that interactions that are faster than a week are erased in an artificial way. If we were to model high-frequency trading news model this is extremely bad since we might be interested in what will happen with the news flow intra-second. In this thesis though we are focused on trying to model the news flow over a longer time period to be able to see if trends in the news flow can be seen for example during economic crisis or a migration stream.

Because of this erasing the fast excitations is probably beneficial for our results since a lot of the excitations otherwise would be very fast and therefore the results we are after would only be visible in a model with more decay parameters, which are very hard to optimise compared to linear parameters. With both computation times and excitation in focus, the choice to bucket together the articles at a suitable resolution seems like a good idea.

With this bucket method, a given time interval $[0, T]$ becomes a grid with intervals of increment size Δt. Thus in this setting, every observation will be on one of the grid points. Likewise, this means that every grid point, or bucket, can store several events. The input time sequence $\mathbf{t}$ is projected on the grid according to

$$\mathbf{t}_G = proj_G(\mathbf{t}), \tag{5.2}$$

where $\mathbf{t}_G$ is the projected time sequence and $proj_G$ is the operator which projects the observed time sequence onto the discretized grid G. Here, the number of events in every bucket in class i and grid point k is denoted by $n_G^{i,k}$. In addition, T_G is denoted as the total number of buckets.

Another thing to note about using exactly 1 week as bucket size is that we limit

the day effects. This is nice because the news flow is considerably lower during the weekends, as was presented in the data chapter, see figure 3.5. This is very bothersome when using a Hawkes model since the excitations does not depend on the weekday. To avoid this some adjustments has to be made but they are either extremely complex or not very good. Because of this, with a normal Hawkes model, the week bucketing will increase our accuracy.

5.2 Hierarchical Structure

In the data set used in the end there are 254 country codes, 56 news groups and 3 sentiments. This makes for $(254 * 56 * 3)^2 = 1820899584$ parameters in the excitation matrix. To get a more stable and robust result decided to decrease the amount of variables. To achieve this two approaches have been taken. First some of the 256 country codes have extremely sparse news reporting. This fact and that some countries should have similar behaviour enabled the clustering of countries close to each other. See table 10.4-10.9 in Appendix A how the regions were defined.

This projection was from 256 dimensions to 34 geography dimensions. These new 34 regions then could be grouped together but they were never split apart. These 34 were the initial grouping of countries in the thesis. No tampering was done with the 56 news groups or the 3 sentiments.

5.3 Time-dependent Background Intensity

Furthermore the news flow is as stated in the data chapter extremely non-stationary. It is increasing in time it seems, at least during the recent years with internet and social media where there are many new platforms to release news. There are also periodic behaviour in most topics, especially the economic news.

This will if left alone create many artificial excitations between classes that share the same type of periodicities or growth but have little to no real correlations between each other. Here the base assumption that the news flow drives itself and that there are no outside factors, such as that seasons exist, falters. The assumption is of course not correct since everything happens in the real world but for an unexpected event the difference is small since that event will spawn articles that exhibit the behaviour of that the first article spawned them. The difference with known outside effects is that the behaviour is uninteresting and the result will be misleading.

This is why it would be optimal to use a background intensity that takes care of all the known outside effects so the excitations are interesting and not used to

model seasonality. This turns out to be not so easy since some of the news topics and countries have very unique patterns. However most news topics have a trend of growing news flow and at least the economic news has a quarter year periodicity. The inhomogenous poisson procees that make up the time-dependent background intensity for the Hawkes Process will therefore be

$$b(t) = b + b_t t + Asin(w_{quarter} t + \phi) \tag{5.3}$$

where b is the constant base intensity, b_t models the linear trend, A the amplitude, $w_{quarter}$ is the quarter year frequency and ϕ the phase shift. All these variables are the estimated for each class in the models in the optimisation step. The phase shift represents that not every country release their quarterly reports at the same time.

5.4 Model 1: Discrete Classes

5.4.1 Spatial Likelihood

From equation 4.27 now the $p(\mathbf{x}|\rho)$ is to be calculated to get the posterior of the parameters to be able to maximise the likelihood. In fact as stated in the theory section, $f_x(\mathbf{x})$ does not depend on the time likelihood why this spatial likelihood part only is needed to compare different class setups. It serves as a penalty to having many article types in 1 class. With this the Bayesian Information Criteria, BIC, can be used to find a good balanced class setup.

Assume that the news are Multinomial distributed. The news types in the classes just projects the news articles from the original class space to the new class space and in that new space get the distribution of old classes in the new classes to get out the probability that an article in a new class is from a certain news type.

$$x^t = P x^t_{old} \tag{5.4}$$

n_x is here the number of articles in the spatial value in the new space, P the projection matrix and x^t_{old} the number of articles in every old class every time step (the bucket implementation discussed in the previous section). Using this the spatial likelihood is just the logarithm of the projection matrix times the amount of old articles in every class

$$\sum_t x^t_{old} \ln \frac{P \circ \sum_t x^t_{old}}{\sum_{\mathcal{X}} P \circ \sum_t x^t_{old}} \tag{5.5}$$

where $\circ$ indicates the element wise product.

5.4.2 Time-dependent Likelihood

The likelihood of the multivariate Hawkes Process can be seen in (4.16). Given the bucketing procedure described in the previous section and the projected sequence of times $\mathbf{t}_G$, the integral term in the log-likelihood of the multivariate Hawkes process stated in equation 4.16 can be written as

$$\int_0^T \lambda_i^*(u)\,\mathrm{d}u = \sum_{k=0}^{T_G} \int_{t_k}^{t_{k+1}} \lambda_i^*(u)\,\mathrm{d}u. \tag{5.6}$$

Consequently, the total log-likelihood of the projected time sequence becomes

$$\log \mathcal{L} = \sum_{i=1}^{n} \left(\sum_{k=0}^{T_G} n_G^{i,k} \log\left(\lambda_i^*(t_k)\right) - \sum_{k=0}^{T_G} \int_{t_k}^{t_{k+1}} \lambda_i^*(u)\,\mathrm{d}u \right) \tag{5.7}$$

Here, it is convenient to write the intensity as a recursive sum. For each i and for all $k \in \{1, \ldots, T_G\}$ it holds that

$$\begin{aligned}
\lambda_i^*(t_k) &= b_i + \sum_{j=1}^{n} \sum_{m=0}^{k-1} V_{ij} n_G^{j,m} e^{-\gamma_j (t_k - t_m)} \\
&= b_i + \sum_{j=1}^{n} V_{ij} \left(n_G^{j,k-1} e^{-\gamma_j (t_k - t_{k-1})} + \sum_{m=0}^{k-2} n_G^{j,m} e^{-\gamma_j (t_k - t_m)} \right) \\
&= b_i + \sum_{j=1}^{n} \left(V_{ij} n_G^{j,k-1} e^{-\gamma_j (t_k - t_{k-1})} + \lambda_{ij}^{*,b}(t_{k-1}) e^{-\gamma_j (t_k - t_{k-1})} \right) \\
&= b_i + \sum_{j=1}^{n} \lambda_{ij}^{*,b,+}(t_{k-1}) e^{-\gamma_j (t_k - t_{k-1})}
\end{aligned} \tag{5.8}$$

For the point where $k = 0$ it holds that $\lambda_i^*(t_0) = \lambda_i^*(0) = b_i$. Here, the exponent label $+$ indicate that it is to the right of the discontinuity. Similarly, the exponent label b indicates that the base intensity has been left out. The last equality comes from the definition that

$$\lambda_{ij}^{*,b,+}(t_k) = V_{ij} n_G^{j,k} + \lambda_{ij}^{*,b}(t_k). \tag{5.9}$$

Because of the time discretisation, the excitation jumps only occur at the grid points. Hence, the integral in equation 5.7 will just be an exponential decay scaled with $\lambda_{ij}^{*,b,+}(t_k)$. It is here obtained that

$$
\begin{aligned}
&\sum_{i=1}^{n}\left(\sum_{k=0}^{T_G} n_G^{i,k}\log\left(\lambda_i^*(t_k)\right) - \sum_{k=0}^{T_G}\int_{t_k}^{t_{k+1}} \lambda_i^*(u)\,\mathrm{d}u\right)\\
&= -\sum_{i=1}^{n} b_i T + \sum_{i=1}^{n}\sum_{k=0}^{T_G} n_G^{i,k}\Bigg(\log\left(\lambda_i^*(t_k)\right)\\
&\quad -\sum_{j=1}^{n}\lambda_{ij}^{*,b,+}(t_k)\int_{t_k}^{t_{k+1}} e^{-\gamma_j(u-t_k)}\,\mathrm{d}u\Bigg) \qquad (5.10)\\
&= -\sum_{i=1}^{n} b_i T + \sum_{i=1}^{n}\sum_{k=0}^{T_G}\left(n_G^{i,k}\log\left(b_i + \sum_{j=1}^{n}\lambda_{ij}^{*,b,+}(t_k)e^{-\gamma_j(t_{k+1}-t_k)}\right)\right.\\
&\quad \left. -\sum_{j=1}^{n}\lambda_{ij}^{*,b,+}(t_k)\frac{1-e^{-\gamma_j\Delta t}}{\gamma_j}\right).
\end{aligned}
$$

This form of the time-dependent likelihood now has a typical recursive form. With this recursive way of describing the likelihood we can construct an algorithm like algorithm 1

Algorithm 1 Distinct Classes Log-likelihood, Explanatory

```
1: procedure LogLike
2:     Project old classes on new class space
3:     Calculate Δt
4:     while t < T do:
5:         Propagate the intensities 1 time step
6:         Integrate intensities 1 time step, subtract from log-likelihood
7:         Take the logarithm of the intensities, add to log likelihood
8:         Add new articles in every class this time step
9:     Sum up log-probabilities for the new classes, add to log-likelihood
10:    return log-likelihood.
```

which is easy enough to understand. Apart from the bucketing version of this algorithm a partly similar algorithm is also proposed in [Bowsher, 2007] and used in [Yang et al., 2017]. However there are ways to make this algorithm even faster than how it is stated above which is beneficial for fast computation.

In the introduction to the model in the theory section it is mentioned that the decay time, γ, is the same for every emitter of excitation. This means that every news article type affects all other news types the same amount of time, but with different amplitude. As a contrast a more complex model could have been used where

γ has an index for both sender and recipient of excitation. This simplification is very convenient for the algorithm since it is possible to propagate the exponentials without taking the linear combination to get out the intensities in every recursive step. The last term above with all sums rearranged for clarity is

$$\sum_{i=1}^{n}\sum_{k=1}^{T_G}\sum_{j=1}^{n}\lambda_{ij}^{*b+}(t_{k-1})\frac{1-e^{-\gamma_j\Delta t}}{\gamma_j}. \tag{5.11}$$

This means that every time step, which goes over the middle sum, it sums over the emitting class j. When the class number becomes large this is very time consuming. The faster version instead is to do the recursive sum first and then sum over the emitting and receiving classes.

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{T_G}\lambda_{ij}^{*b+}(t_{k-1})\frac{1-e^{-\gamma_j\Delta t}}{\gamma_j}. \tag{5.12}$$

Now recall that the only i-dependence in $\lambda_{ij}^{*b+}(t_{k-1})$ is in V_{ij}. Now take out the factor V_{ij} and denote that $\lambda_j^{*bV+}(t_{k-1})$ and the term becomes

$$\sum_{i=1}^{n}\sum_{j=1}^{n}V_{ij}\sum_{k=1}^{T_G}\lambda_j^{*bV+}(t_{k-1})\frac{1-e^{-\gamma_j\Delta t}}{\gamma_j}. \tag{5.13}$$

which is suited for parallel computations. With this we can construct an algorithm to calculate the log-likelihood which is faster than the original setup. Using the notation that a variable without index i or j is the respective matrix or vector it becomes

$$\sum_{i=1}^{n}V\sum_{k=1}^{T_G}\lambda^{*bV+}(t_{k-1})\frac{1-e^{-\gamma\Delta t}}{\gamma}, \tag{5.14}$$

and with this the algorithm can be constructed which can be seen in algorithm 2

Algorithm 2 Discrete Classes Log-likelihood, Mathematical

1: **procedure** LOGLIKE
2: $\mathbf{t}_G \leftarrow Proj_G(\mathbf{t})$
3: $\Delta t \leftarrow \frac{T}{T_G}$
4: $\boldsymbol{I}_{\Delta t} \leftarrow \frac{1-e^{-\gamma \Delta t}}{\gamma}$
5: $t = -1$
6: $\boldsymbol{\lambda}_{-1}^{*,b,V,+} \leftarrow 0$
7: **while** $t < T_G$ **do**:
8: $\boldsymbol{\lambda}_{t+1}^{*,V,b} \leftarrow \boldsymbol{\lambda}_{t}^{*,b,V,+} e^{-\gamma \Delta t}$
9: $\boldsymbol{l}_{log,t+1} \leftarrow \boldsymbol{n}_G^{t+1} \log\left(\boldsymbol{b} + \boldsymbol{V}\boldsymbol{\lambda}_{t+1}^{*,V,b}\right)$
10: $\boldsymbol{\lambda}_{t+1}^{*,b,V,+} \leftarrow \boldsymbol{\lambda}_{t+1}^{*,b,V} + \boldsymbol{n}_G^{t+1}$
11: $t \leftarrow t + 1$
12: $\log p(\mathbf{t}|b, V, \gamma) \leftarrow \text{sum}\left(-\boldsymbol{b}T + \sum_{t=0}^{T_G} \boldsymbol{l}_{log,t} - \boldsymbol{V} \sum_{t=0}^{T_G} (\boldsymbol{\lambda}_t^{*,b,V} I_{\Delta t})\right)$
13: Calculate $\log p(\mathbf{x}|\rho)$
14: $\log p(\mathbf{y}|b, V, \gamma, \rho) \leftarrow \log p(\mathbf{t}|b, V, \gamma) + \log p(\mathbf{x}|\rho)$
15: **return** $\log p(\mathbf{y}|b, V, \gamma, \rho)$.

Note here row 9 in the algorithm which is the big time consumer. This is because the logarithm is not a linear operator and hence taking the sum over t first is not possible like (5.14) showed was possible for the integral part of the likelihood. This would be a big increase in performance if any fast approximations to calculate this could be made, but this has not been tested in this thesis.

5.5 Overlapping Classes

When the classes are overlapping the likelihood function can no longer be split up in time and space. This is because the excitation kernel will be dependent on both variables whereas before it was only dependent on time. As in the time dependent part of the discrete classes it is convenient to start with the log-likelihood which with the same bucketing and hierarchical structure as before

$$\log \mathcal{L} = \sum_{i=1}^{n} \left(\sum_{k=0}^{T_G} n_G^{x_k,k} \log\left(\lambda_i^*(t_k, x_k)\right) - \sum_{k=0}^{T_G} \int_{t_k}^{t_{k+1}} \int_{x \in \mathcal{X}} \lambda_i^*(u, x) dx du \right) \tag{5.15}$$

The trouble for the computations is now that the intensity no longer is just the intensity of a categorical variable in time but an intensity in space and time. The recursive behaviour needed to speed up everything now has to include the spacial

variable and then

$$
\begin{aligned}
&\lambda(t_k, x_k) = b(x_k) + \sum_{m=0}^{k-1} \nu(t - t_m, x, x_m) = \\
&= \sum_{i=1}^{N} b_i g^i(x_k) + \sum_{m=0}^{k-1} \sum_{i=1}^{N} \sum_{j=1}^{N} V_{ij} g^i(x_k) \frac{g^j(x_m)}{\sum_{l=1}^{N} g^l(x_m)} n_G^{x_m, m} e^{-\gamma_j (t_k - t_m)} = \\
&= \sum_{i=1}^{N} b_i g^i(x_k) + \sum_{i=1}^{N} \sum_{j=1}^{N} V_{ij} g^i(x_k) \Bigg(\frac{g^j(x_{k-1})}{\sum_{l=1}^{N} g^l(x_{k-1})} n_G^{x_{k-1}, k-1} e^{-\gamma_j (t_k - t_{k-1})} + \\
&+ \sum_{m=0}^{k-2} \frac{g^j(x_m)}{\sum_{l=1}^{N} g^l(x_m)} n_G^{x_m, t_m} e^{-\gamma_j (t_k - t_m)} \Bigg) = \\
&= \sum_{i=1}^{N} b_i g^i(x_k) + \sum_{i=1}^{N} \sum_{j=1}^{N} \Bigg(V_{ij} g^i(x_k) \frac{g^j(x_{k-1})}{\sum_{l=1}^{N} g^l(x_{k-1})} n_G^{x_{k-1}, k-1} e^{-\gamma_j (t_k - t_{k-1})} + \\
&+ \lambda_{ij}^{*b}(t_{k-1}, x_{k-1}) e^{-\gamma_j (t_k - t_{k-1})} \Bigg) = \\
&= \sum_{i=1}^{N} \left(b_i g^i(x_k) + \sum_{j=1}^{N} \lambda_{ij}^{*b+}(t_{k-1}, x_{k-1}) e^{-\gamma_j (t_k - t_{k-1})} \right)
\end{aligned}
\tag{5.16}
$$

where just as in the discrete case the last equality is just a simplification of the terms above with

$$
\lambda_{ij}^{*b+}(t_{k-1}, x_{k-1}) = V_{ij} g^i(x_k) \frac{g^j(x_{k-1})}{\sum_{l=1}^{N} g^l(x_{k-1})} n_G^{x_{k-1}, k-1} + \lambda_{ij}^{*b}(t_{k-1}, x_{k-1}). \tag{5.17}
$$

Before the expression is inserted into the log likelihood the reader should be reminded that $g^i(x)$ is the probability density function of a class and therefore by definition

$$
\int_{x \in \mathcal{X}} g^i(x) = 1. \tag{5.18}
$$

Now inserting the recursive expression for the intensity into the log likelihood

$$
\begin{aligned}
&\sum_{k=0}^{T_G} n_G^{x_k,k} \log\left(\lambda^*(t_k, x_k)\right) - \sum_{k=0}^{T_G} \int_{t_k}^{t_{k+1}} \int_{x\in\mathcal{X}} \lambda^*(u,x) dx du = \\
&= \sum_{k=0}^{T_G} n_G^{x_k,k} \log\left(\sum_{i=1}^{N}\left(b_i g^i(x) + \sum_{j=1}^{N} \lambda_{ij}^{*b+}(t_{k-1}, x_{k-1}) e^{-\gamma_j(t_k - t_{k-1})}\right)\right) - \\
&- \sum_{k=0}^{T_G} \int_{t_k}^{t_{k+1}} \int_{x\in\mathcal{X}} \sum_{i=1}^{N}\left(b_i g^i(x) + \sum_{j=1}^{N} \lambda_{ij}^{*b+}(t_{k-1}, x_{k-1}) e^{-\gamma_j(t_k - t_{k-1})}\right) dx du = \\
&= \sum_{k=0}^{T_G} n_G^{x_k,k} \log\left(\sum_{i=1}^{N}\left(b_i g^i(x) + \sum_{j=1}^{N} \lambda_{ij}^{*b+}(t_{k-1}, x_{k-1}) e^{-\gamma_j(t_k - t_{k-1})}\right)\right) - \\
&- \sum_{k=0}^{T_G} \int_{t_k}^{t_{k+1}} \sum_{i=1}^{N}\left(b_i + \sum_{j=1}^{N} \lambda_{ij}^{*bg+}(t_{k-1}, x_{k-1}) e^{-\gamma_j(t_k - t_{k-1})}\right) dx du = \\
&= \sum_{k=0}^{T_G} n_G^{x_k,k} \log\left(\sum_{i=1}^{N}\left(b_i g^i(x) + \sum_{j=1}^{N} \lambda_{ij}^{*b+}(t_{k-1}, x_{k-1}) e^{-\gamma_j(t_k - t_{k-1})}\right)\right) - \\
&- \sum_{i=1}^{N} b_i T - \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=0}^{T_G} \lambda_{ij}^{*bg+}(t_{k-1}, x_{k-1}) \frac{1 - e^{-\gamma_j \Delta t}}{\gamma_j}.
\end{aligned}
\tag{5.19}
$$

This log likelihood has great similarities with the one for the discrete classes case. This version includes some additional computational difficulties from the density functions which are not present in the discrete case.

The algorithm used in Tensorflow to estimate the parameters using this can be seen in algorithm 3. Because of the similarities with the discrete case the explanatory version of the algorithm is omitted since the differences are better highlighted in a detailed framework.

Algorithm 3 Overlapping Classes log-likelihood, Mathematical

procedure LogLike
  $\mathbf{g}(\mathbf{x}) \leftarrow \mathbf{g}(\mathbf{x})$
  $\mathbf{d}(\mathbf{x}) \leftarrow \frac{\mathbf{g}(\mathbf{x})}{\sum_{l=1}^{N} g^l(\mathbf{x})}$
  $\boldsymbol{I}_{\Delta t} \leftarrow \frac{1-e^{-\gamma \Delta t}}{\gamma}$
  $\boldsymbol{b}(\mathbf{x}) \leftarrow \sum_{i=1}^{N} b_i g^i(\mathbf{x})$
  while $t < T$ **do**:
    $\boldsymbol{\lambda}_{t+1}^{Vb} \leftarrow \boldsymbol{\lambda}_{t}^{Vb+} e^{-\gamma \Delta t}$
    $\boldsymbol{\lambda}(\mathbf{x}, k) \leftarrow \boldsymbol{b}(\mathbf{x}) + \sum_{i=1}^{N} \mathbf{g}(\mathbf{x}) \mathbf{V}. * \lambda_{t+1}^{Vb}$
    $\mathbf{l}_{log,t+1} \leftarrow \mathbf{n}_G^{\mathbf{x},k} \log\left(\boldsymbol{\lambda}(\mathbf{x}, k)\right)$
    $\boldsymbol{\lambda}_{t+1}^{Vb+} \leftarrow \lambda_{t+1}^{Vb} + \mathbf{d}(\mathbf{x}) n_{\mathbf{x}}^{t+1}$
  $\ln \text{Likelihood} \leftarrow \sum_{i=0}^{N} (\boldsymbol{V} \sum_{t=0}^{T} (-\boldsymbol{\lambda}_t^{Vb} \boldsymbol{I}_{\Delta t}) - \boldsymbol{b} T) + \sum_{t=0}^{T} \boldsymbol{l}_{log,t}$
  return $\ln$ Likelihood.

Chapter 6

Methodology

The way this study was set up two new methods to model the flow of news and sentiment was tried out. To get a base line the Poisson model was used. This comparison is because the Hawkes process is a Poisson process with added complexity in terms of the excitation kernel.

For the different models is important to get results which are interpretable and comparable. This is to be able to tell which model is actually beneficial to use. Also it is needed for the reader to be able to grasp the results from the study. To be able to achieve both these this thesis was conducted testing a few different cases.

6.1 Initialisation and Prior, Classes

How to divide the classes is a very difficult question in this thesis which we are not to delve too deep into. We are interested in what countries should be in the same groups to see maximal excitations. The trouble here is that discrete classes are not differentiable and class optimisation for a Poisson model makes little to no sense. In the overlapping model the classes are optimisation is possible, however the problem is not convex and what initialisation is used is important for the end result. This will not be examined too much but more give a comparison to the distinct case.

There is also another complexity which is time consumption, data and number of variables. The natural starting point and base line would obviously be to test the classes which Ravenpack is using in their hierarchy, all countries separately and the sentiments separately. However as discussed above the 56 news topics and 34 regions will here be treated as the ground truth, the most complex case.

There are 56 different topics in the Ravenpack data. The first tests will be done with all of them in different groups. We also want to see if there is a merit in using the overlapping classes and for that we are mostly interested in projecting the 56

groups on a lower dimensional space. We will therefore limit the number of groups with a prior of 5 to see if the amount of groups are vastly exaggerated in Ravenpack. To initialise the overlapping case the multinominal distribution with just random initialisation will be done.

In the second layer of the hierarchy we will try 3 different initialisations/prior combinations. In the first case the 34 regions described in section 5.2 will be used. The second one will be all regions in the same group. This is the same as saying that the news flow does not depend on where the article comes from but only the content of it. Lastly like for the topics we will have 5 groups for the overlapping model to see if 34 regions are needed at all.

For the sentiment all the three groups negative, neutral and positive news will all be used in every test to see if it is possible to see if there is a generally positive or negative news flow right now in a certain group.

6.2 Data Set

In the data section there are some important things to consider when choosing the final data set. The bucketing mentioned in the model section directly affects the behaviour of the model. There are however some cleaning of the data that can be done to get more meaningful results. In this study 2 main problem with data were found out.

Firstly some topic-region combination were very rare in the news flow. An example of that is War related news about Antarctica. Out of the 5712 topic-region-sentiment combinations in the thesis 2449 were happening in less than 100 weeks out of the 521 weeks ranging from January 2007 to December 2016. This is a huge problem in numerical optimisation since it is very likely that there can be found a connection between to classes by coincidence. In figure 6.1

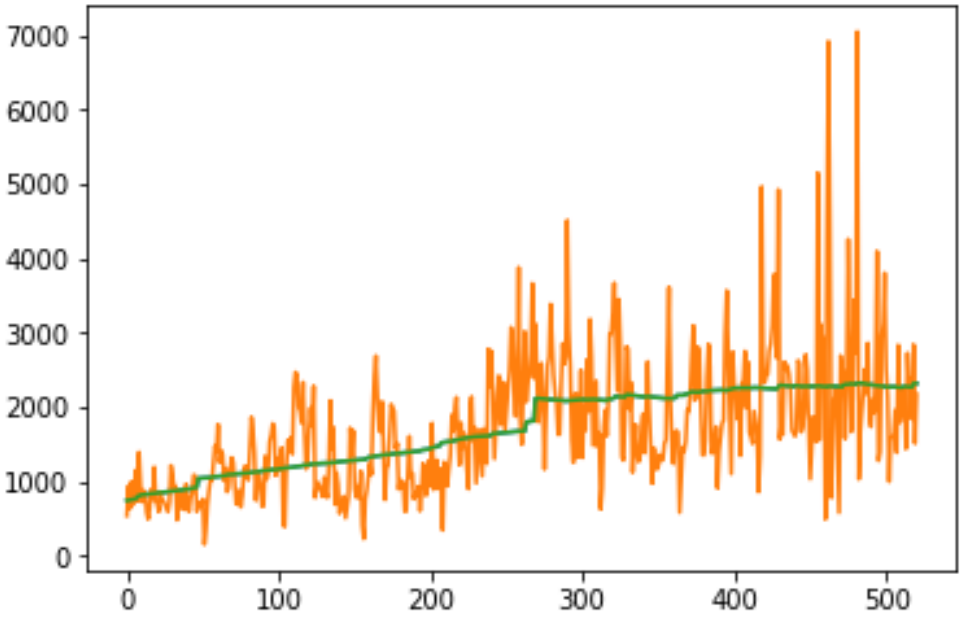


Figure 6.1: Number of articles per week in group Negative news in Europe in orange. The green is the expected number of observations in the group. Notice the weird jump at week 270 from the group Negative news in Antarctica.

there is a test without the cleaned data set where a single observation of news in Antarctica at week 270 affects the news flow of news in Europe. This excitation is magnitudes bigger than the other excitations in this early test and it would mean that Europe is closest connected to Antarctica when it comes to news flow. Because of this the data set used in the actual tests had all these 2449 combinations with very few observations taken away.

The other big characteristic that some groups had were that they were extremely periodic. These groups were big economic groups and even though there probably are some information in the news flow our models were unable to detect it. Instead the only behaviour that was caught was the periodic behaviour using whatever excitation it could to model this, or in the time-dependent model there were no excitation at all, only a sinusoid with the correct amplitude. in figure 6.2 4 groups are shown.

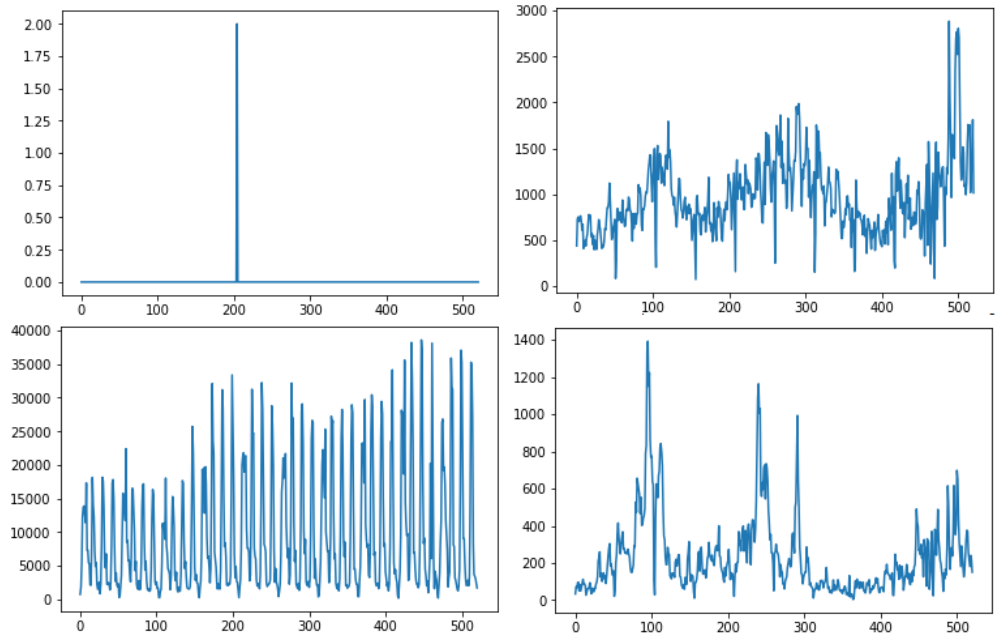


Figure 6.2: To the left, two types of classes that are taken away from the study. In the top left corner too few observations and in the bottom right too periodical news flow. To the right are two groups that are kept.

The groups to the left are examples of groups that were take away because of too few observations or too periodic news flow. To the right in the figure are examples of groups that were kept. The periodic topics that were taken away were Earnings, Dividends and revenues. They were taken away for all regions and sentiments. The small groups that were taken away were more random since some topics only contain positive news and vice versa and some countries contain less news about a specific topic. The condition was as stated above though, observations in more than 100 weeks over the 10 years and the group was kept.

The 3 periodic topics consist of 918 total region-topic-sentiment combinations which were also taken away from the data set. The final data set hence consisted of $5712 - 2449 - 918 = 2345$ groups in total. In the top right figure one can also see the Christmas week dips every year which are not handled in any way. As with the argument about using week buckets instead of days where some days had lower flow the same is true for the Christmas week. To solve this a more complex model or yearly buckets would be needed.

6.3 Largest Excitation Model Choice

The structure of the multi-variate Hawkes process with a large amount of classes is one that contain a massive amount of excitation parameters. This is a classic problem that might lead to overtraining and will therefore get less predictive power. Even though the likelihood of the training set is high it does not necessarily mean that the parameters are well tuned for the future. For this the BIC is used and this measure will be better if the amount of variables decreases.

Therefore the lasso model selection can be used where variables are forced towards 0 and when an appropriate amount of parameters are 0 the non-zero elements will be re-estimated. In the cases that we are interested in it seems that most of the parameters that are uninteresting are already estimated to almost 0 when not using penalty since the model notices that there are no excitation. Therefore a slightly simpler version will be used where if $V_{ij} \sum_{t=0}^{T_G} n_G^t$ is bigger than a certain threshold the parameter is kept and otherwise made constant 0. This threshold is chosen so the visible excitations are kept while the excitation that are not noticeable are thrown away.

6.4 Likelihood and Results

In this thesis there is a real question what results should actually be presented. Displaying all optimised values will not be possible simply because of how many they are.

The most obvious result that will be presented is of course which models that are the most likely. This will be related to the amount of variables with the BIC which can give an indication on what model that should be used.

6.4.1 Excitation

The main result for this thesis though is to show how different topics and regions are related in the news flow. This is hidden in the excitation matrix of the models. Because of the sheer size of this matrix though it is not possible to extract any meaningful information by presenting the matrix itself. Instead a node graph will be displayed that will function as a world map of how closely related different groups are to each other.

Specific excitation results from interesting groups will also be presented individually. This is to get a qualitative feeling on how intuitive the excitations found in the thesis are. These results should of course be read with caution since the classes shown in the report will have been hand picked to show the interesting part of them.

6.4.2 Intensity

The excitations are a paramter that in themeselves have interpretation value in the sense that they model dependence. The main result of the models however are the intensities of the different groups. Getting a feeling for how good the model is at capturing the behaviour of the classes will be much easier by watching the intensity graph instead of trying to figure out what a good likelihood should be.

The maximum likelihood estimate can then be tried out on test data by using the estimated excitations and plotting the intensity of the test year. This will also be done and will provide the reader material to determine if the models are overfitted to the training set or how well it performs on future data.

The last thing that will be presented in this thesis is the performance of the model during some of the extreme events during the last few years. This will be a zoom in version of the test above where Brexit and the Arab Spring will be examined. This will by no means be a quantitative part of the study, but a part where it is possible to understand the models limits and merits in a crisis.

Chapter 7

Results

This now brings us to the result part. In this section there will be a variety of Hawkes process and Poisson process groups compared to each other as explained in the previous chapter. To compare the different groups the aforementioned the Bayesian Information Criteria and Likelihood will be used. To get a deeper understanding about how groups affect each other the excitations will be studied in detail as well.

7.1 Group Selection A: 53 News Types, 1 Geography Section, 3 Sentiments

For the first 2 group selections, A and B, the focus is to see whether there are some dependencies that are apparent in only one dimension, either in the News space in A or the Geography space in B. For A the negative log likelihoods can be found in table 1

Table 7.1: Likelihoods of different models. There are 171 728 259 observations.

Model	Log Likelihood	Parameters	BIC
Homogeneous Poisson	-617 285 166	2 345	1 234 614 797
Homogeneous Hawkes	-608 963 096	30 737	1 218 509 009
Time-dependent Poisson	-609 285 099	2 849	1 218 624 219
Time-dependent Hawkes	-606 954 583	31 241	1 214 501 540
Reduced T-d. Hawkes	-606 949 327	4 658	1 213 986 976

The BIC indicates that the reduced version of the Hawkes process seems to be the best. This is reasonable since the Hawkes model is more advanced than the Poisson process and with the addition of variable reduction the amount of parameters decreases to a point where the important excitations shine through. In figure 7.1 the different models are compared. The homogeneous poisson is left out since it is just a straight line in the middle.

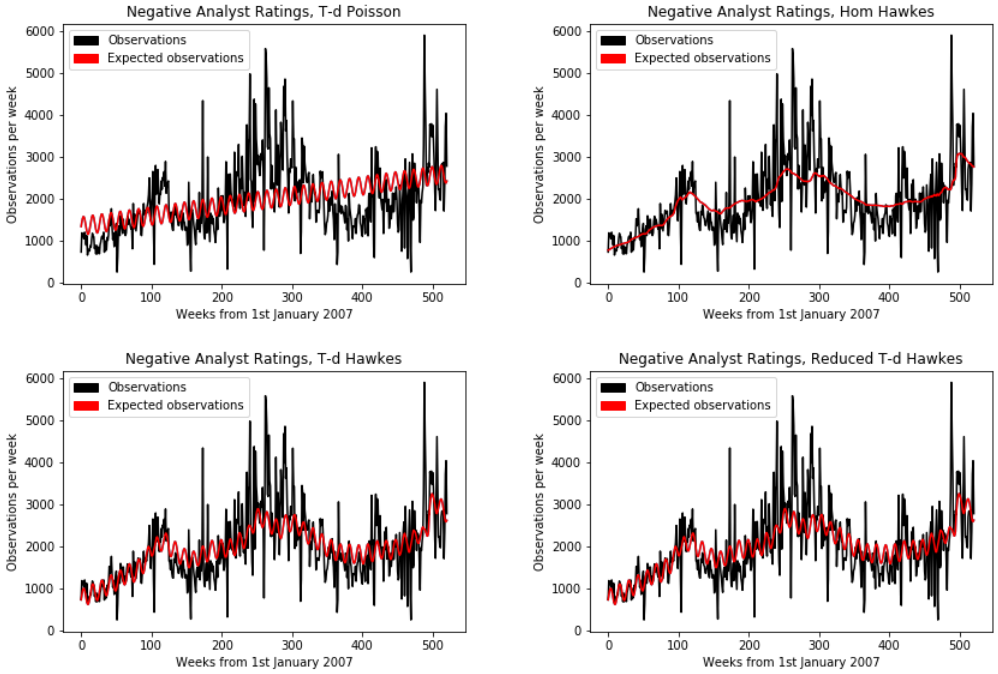


Figure 7.1: The group Negative Analyst Ratings plotted against the expected number of articles of the different models.

In the top left image we can see that it is beneficial for the model to have some seasonality. Also a growing trend seems preferable, at least when not having excitations. The seasonsality seems to be kept when using the Hawkes models at the bottom. We can not the Brexit spike in the end that will be seen more further down.

7.1.1 Excitation Between Classes

The excitation between the news groups placing all regions in the same bucket can be seen in figure 7.2. In the figure an algorithm called Force Atlas 2 [Jacomy et al., 2014] is used. This work as a gravitational model where more excitations attract to a larger degree. This will cluster together groups that are closely related. We want in this first example show the connections between economic news and news related to something else.

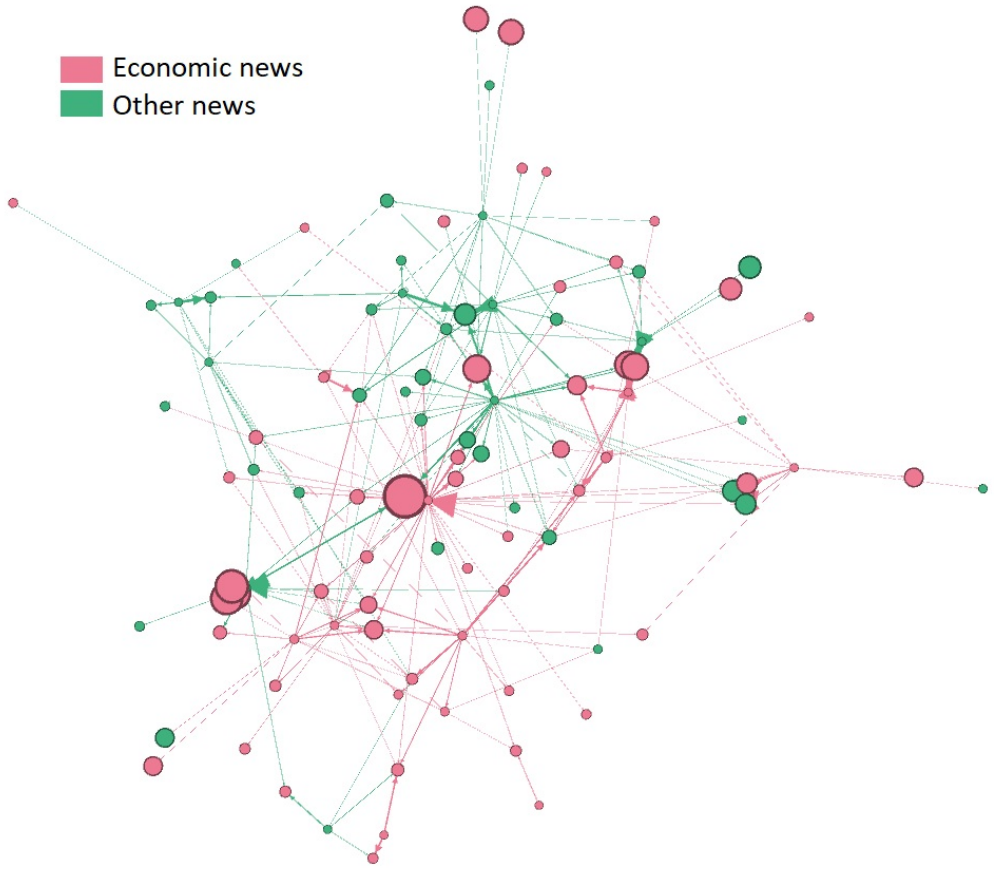


Figure 7.2: A visualisation of the dependencies of the Reduced Time-dependent Hawkes model between the news groups. The green nodes and arrows represent the economic news of the news flow and excitations from them. The red nodes and arrows are the counterpart for the news not directly related to economic reporting. The size of the group is represented by the size of the node and the size of the excitation by the size of the arrow.

In the figure big nodes corresponds to classes with a large amount of news and the size of the arrows corresponds to the size of the excitation of between the classes. In the figure on can see that most of the economic news are in the bottom right corner. In the green part of the graph the biggest node is negative war and conflict. The nest in the green part that excites a lot of groups is the negative aid group. Other groups around those nests seem to be groups related to the world in general with war, civil unrest and natural disasters among them. In the red economic corner the only big influencer seem to be negative price targets. From it is Bankruptcy, Credit ratings and alike excited.

To get a better understanding of specific excitations between the classes the reader can see the excitations from some of the nodes zoomed in in the following figures. In figure 7.3 the big green node Negative War and Conflict is highlighted.

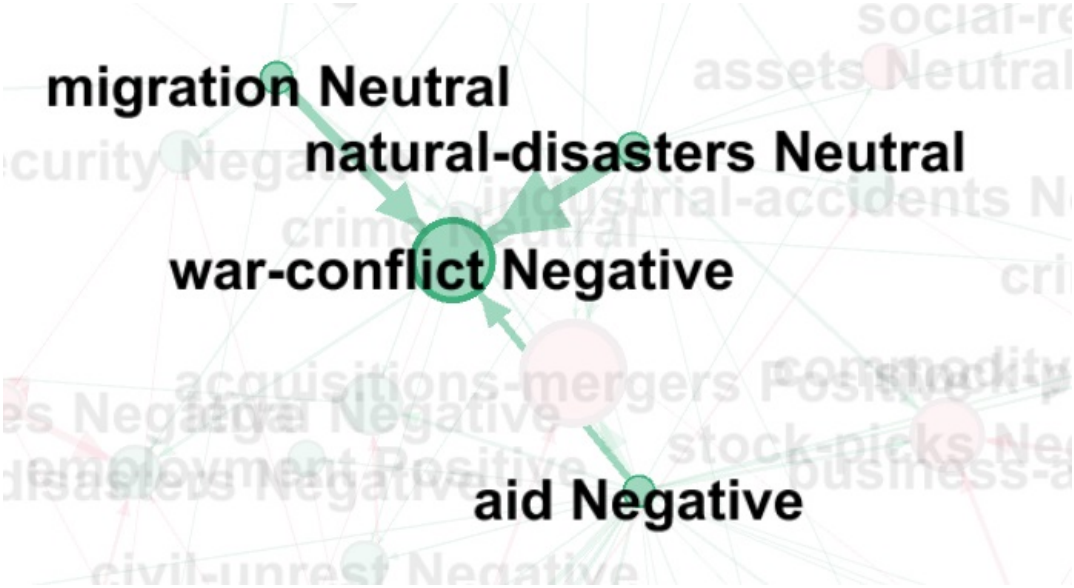


Figure 7.3: A visualisation of the dependencies between the news group Negative War and Conflict and its dependent groups.

As can be expected migration and aid seems to be related to the war group. Natural disasters may not be as intuitive. From the results it seems though that an unstable weather highly affects the conflicts in the world. The excitation is quite strong and it seems unlikely that this should be a coincidence, but it is a possibility. Moving on to the economic part of the graph there is an interesting result in the group Negative Domestic Product, figure 7.4.

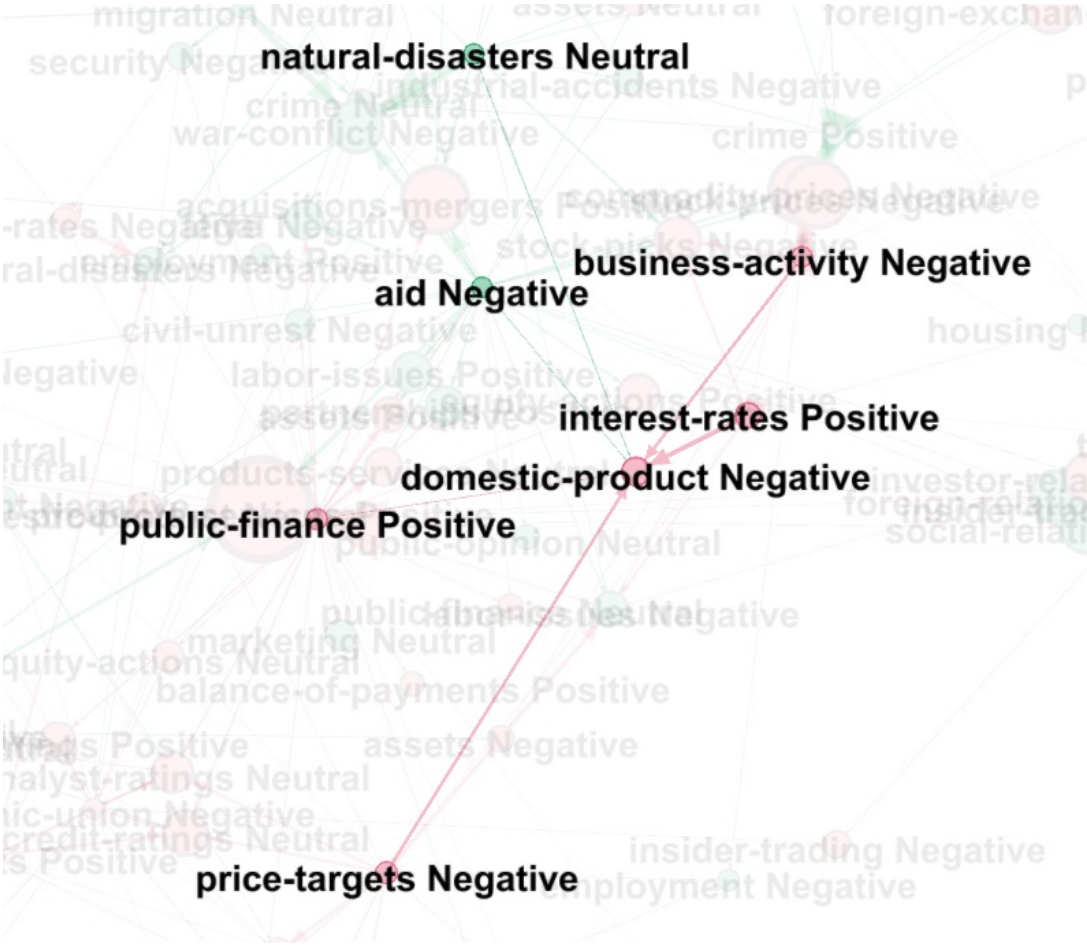


Figure 7.4: A visualisation of the dependencies between the news group Negative Domestic Product and its dependent groups.

The most prominent excitation is from positive interest rates. Having some caution towards the direction of the arrow it seems that a weak domestic product is related to interest rates changes that will affect the market in a positive way. This has been seen time and time again the last few years when the governments are trying to manoeuvre the tough financial times that have been. The big influencer in the economic news flow was though as stated the negative price targets, for that node see figure 7.5.

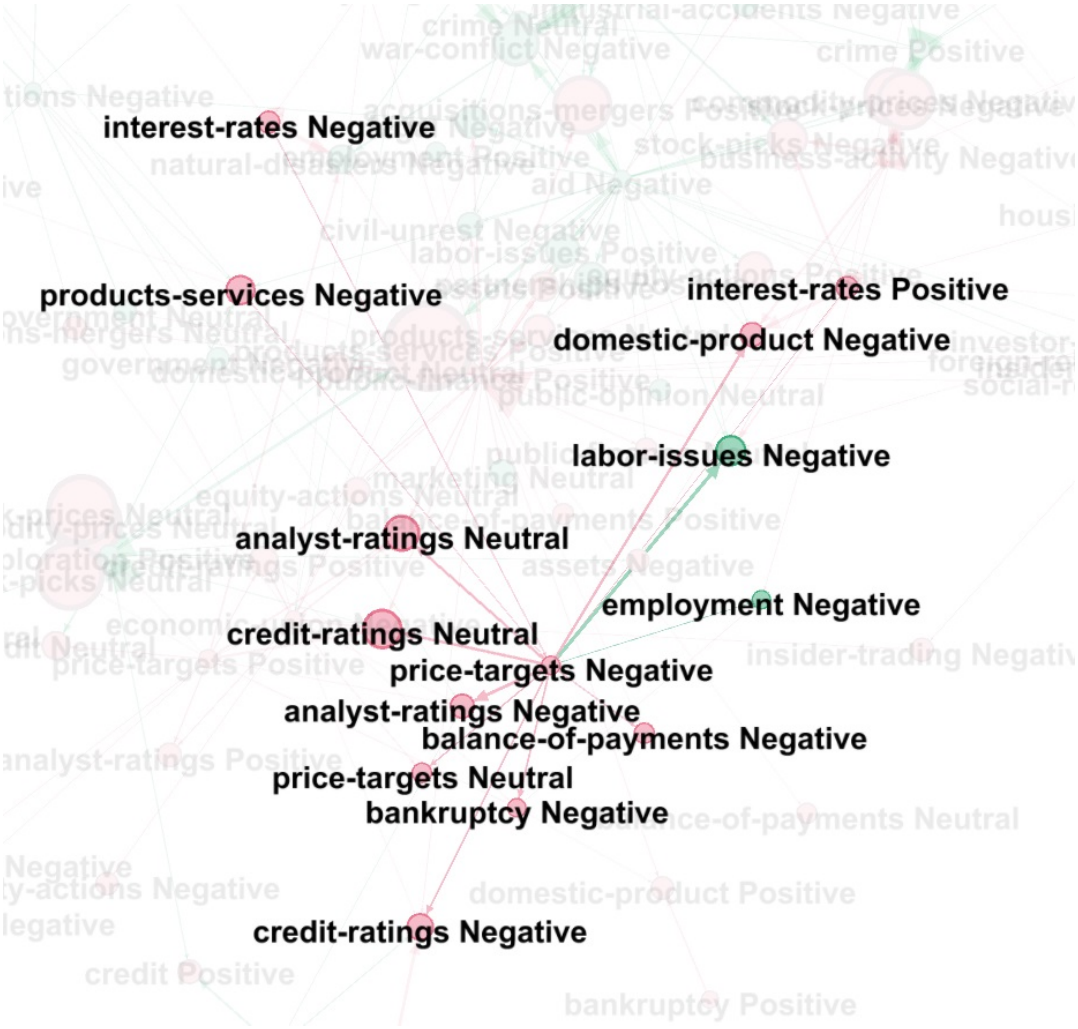


Figure 7.5: A visualisation of the dependencies between the news group Negative Price Targets and its dependent groups

That the price targets are the origin for bankruptcy is quite hard to imagine. Again we have to consider the possibility that the arrow direction is not correct. In section 7.3.2 there is a discussion whether the size of the groups affect the direction of the excitation too much. If that is the case is hard to know from our results. Of course the result can be correct since the news flow could start report that the financial times are tough before many companies declare bankruptcy.

The other groups excited from the negative price targets seem to almost exclusively be negative groups. It is very reasonable with how one would think it should be. Employment and labour issues are the only groups that are not directly eco-

nomic groups that are affected. Going deeper in the analysis it can be noted that the three groups that are connected to employment are negative bankruptcy, negative business activity and negative price targets. This could be valuable information to catch an increasing unemployment rate just by watching the news flow.

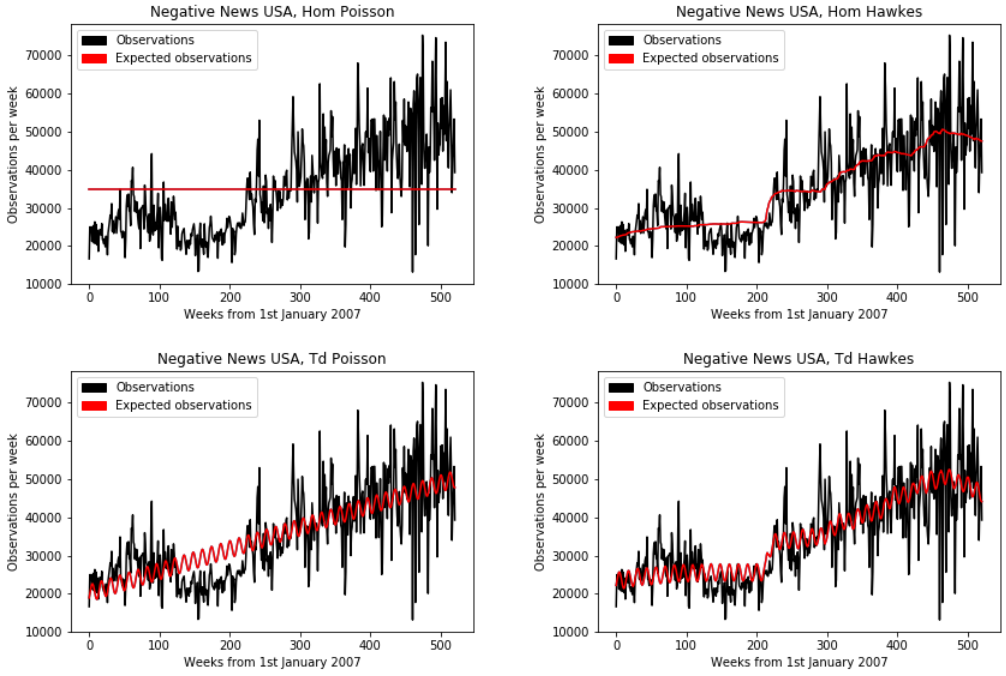
7.2 Group Selection B: 1 News Type, 34 Geography Sections, 3 Sentiments

In table 7.2 the likelihoods and BIC is shown when only splitting the data set over regions. All news topics are placed in the same bucket and the 34 regions and 3 sentiments make up the groups.

Table 7.2: Likelihoods of different models. There are 171 728 259 observations.

Model	Log Likelihood	Parameters	BIC
Homogeneous Poisson	-617 285 166	2 345	1 234 614 797
Homogeneous Hawkes	-609 648 087	12 851	1 219 539 847
Time-dependent Poisson	-610 375 511	2 651	1 220 801 289
Time-dependent Hawkes	-608 700 453	13 157	1 217 650 381
Reduced T-d. Hawkes	-608 697 141	3 811	1 217 466 544

First we note that the likelihoods in general are higher than when dividing the groups in topics. We note that the amount of groups are fewer though so the BIC may be a better measure. These are however also worse than their topic groups counterpart. We can also seen in figure 7.6 that the fit seems to be worse than in the other example.



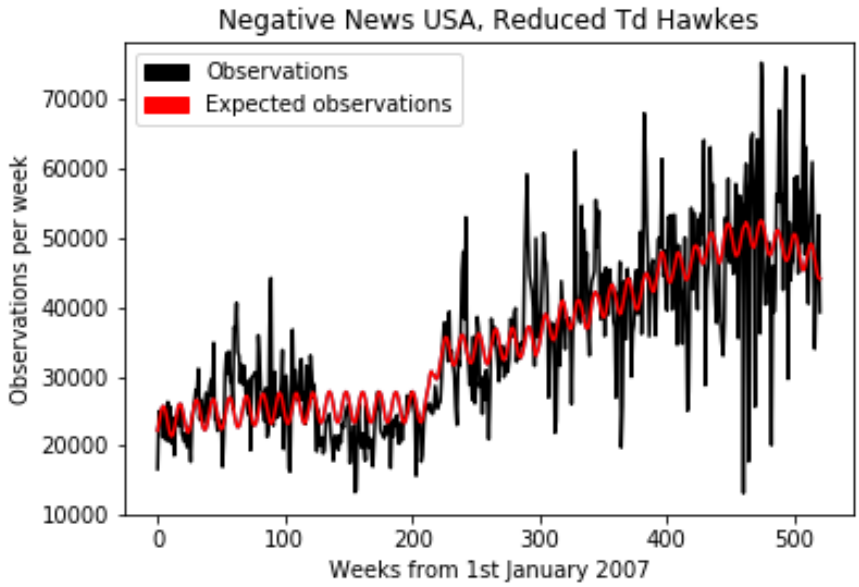


Figure 7.6: The group Negative news in USA plotted against the expected number of articles of the different models.

To compare random groups is definitely not a perfect way to evaluate a model but the figures are meant to show the reader whats captured in the respective model.

For example the peak in the beginning and the dip at week 150 are completely missed by the model. When examining the node map, the knowledge of that the model might not be perfect is good to have.

7.2.1 Excitation Between Classes

To see what regions that affect each other we will define larger regions than the 34 where regions that are close to each other make up a bigger new region. This is to see if regions close to each other will be close to each other in the map. The division of these new regions can be seen in figure 7.7. Each colour in the map will then represent the area in the node map in figure 7.8.

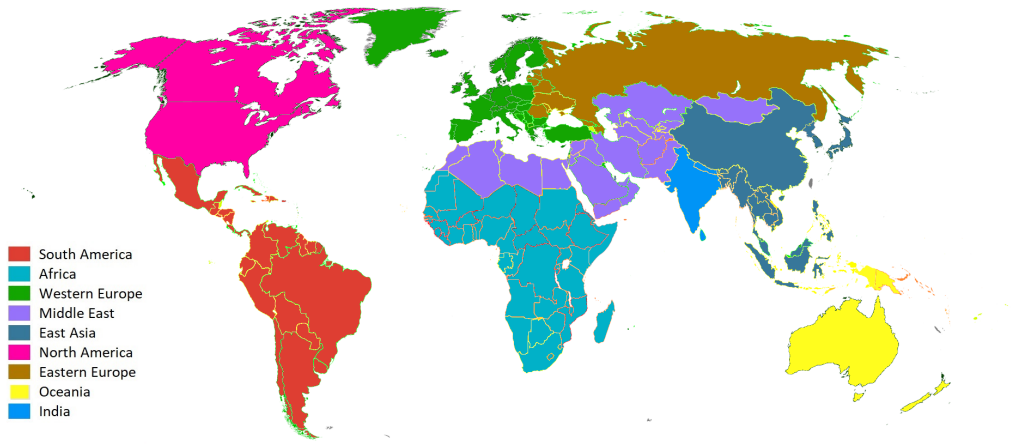


Figure 7.7: A map over the new bigger regions that will have their own colour in the following node maps. The colour in this map corresponds to the colour in the graphs following.

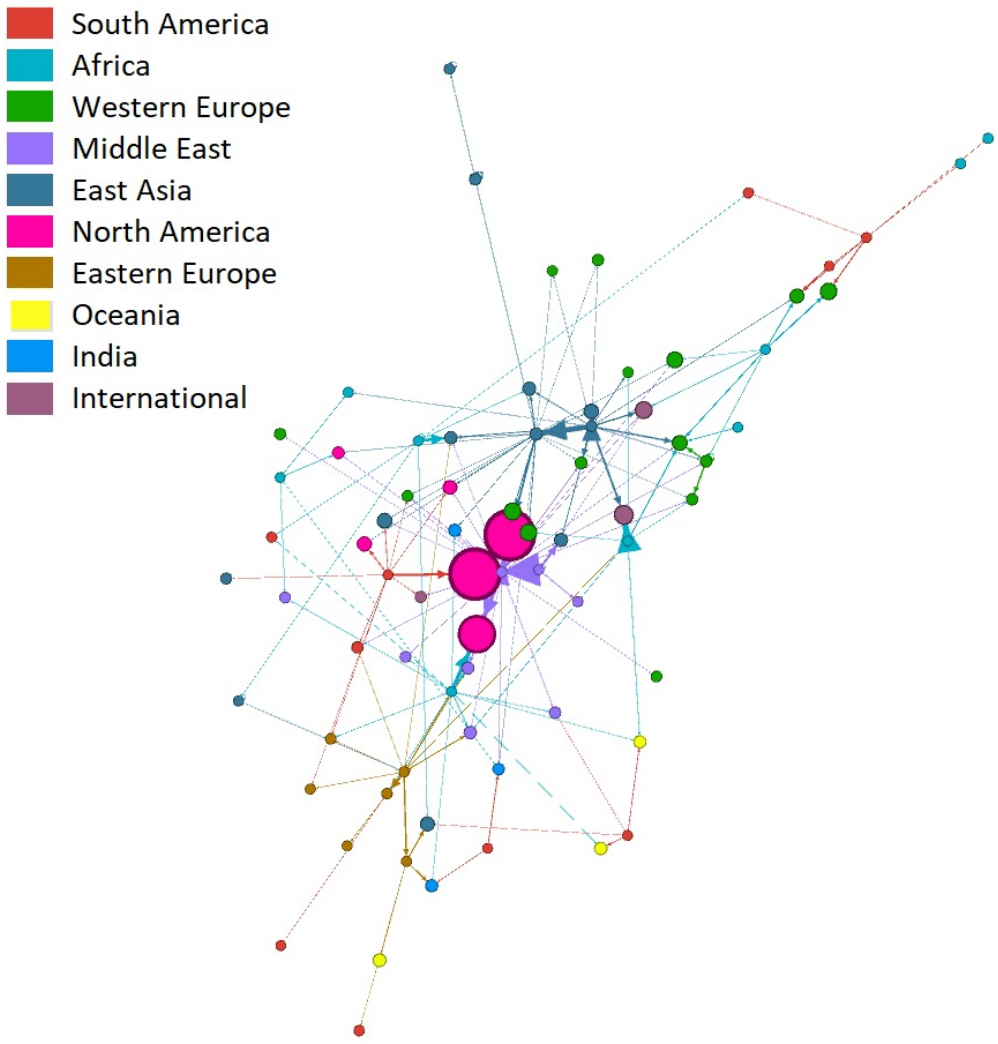


Figure 7.8: A visualisation of the dependencies between the news groups for the 155 most important excitations of the Reduced Time-dependent Hawkes model. Each big region is represented by a unique colour. The size of the group is represented by the size of the node and the size of the excitation by the size of the arrow.

In the map the North American news flow dominate in terms of amount of news and many groups are exciting that region. It seems that North America does not excite any other country though. A cluster of the larger group Eastern Europe can be seen in the bottom left part with the groups Negative Russia and Negative, small region, Eastern Europe in the center. Also the East Asia cluster in then top seems to have noticeable excitations between each other. All different regions in that larger group are represented with big excitations to or from China, Japan/Korea, Indochina and

South East Asia. Lastly we notice the large purple arrows exciting news from the Middle East big group, including North Africa, affecting the North American news flow in a very substantial way. This will be seen more in detail in the next section, with the more complex model with both regions and topics.

7.3 Group Selection C: 53 News Types, 34 Geography Sections, 3 Sentiments

The results in each dimension indicates that there the dependencies in both topics and region are the strongest in regions that are close to each other and in topics that by intuition seems to be closely related. This brings us to the results that takes into account both topics and regions at the same time. In table 7.3 the likelihoods and BIC for the test with 2345 total classes can be seen.

Table 7.3: Likelihoods of different models. There are 171 728 259 observations.

Model	Log Likelihood	Parameters	BIC
Homogeneous Poisson	-617 285 166	2 345	1 234 614 797
Homogeneous Hawkes	-597 915 193	5 503 715	1 300 188 659
Time-dependent Poisson	-607 324 146	9 380	1 214 826 150
Time-dependent Hawkes	-596 481 879	5 510 750	1 297 455 425
Reduced T-d. Hawkes	-595 931 403	84 966	1 193 473 882

As expected if the optimisation found a global minimum the likelihood for the most complex case is better than for the simpler cases. Also if using the reduced model choice the BIC indicate that predictions when splitting up regions and topics are more powerful.

When looking at the expected observations every week in figure 7.9 one can see extreme behaviour at week 490. This is Brexit and the actual group exciting is Negative Economic Union in Great Britain. This is a problem even with our reduced model. It has chosen the excitations that are most important, but with foreknowledge of what has happened. This is something that would need more investigation if using the model in prediction purpose. In this thesis we are primarily interested in finding the connections in the news world in the recent years so we will not delve too deep into this.

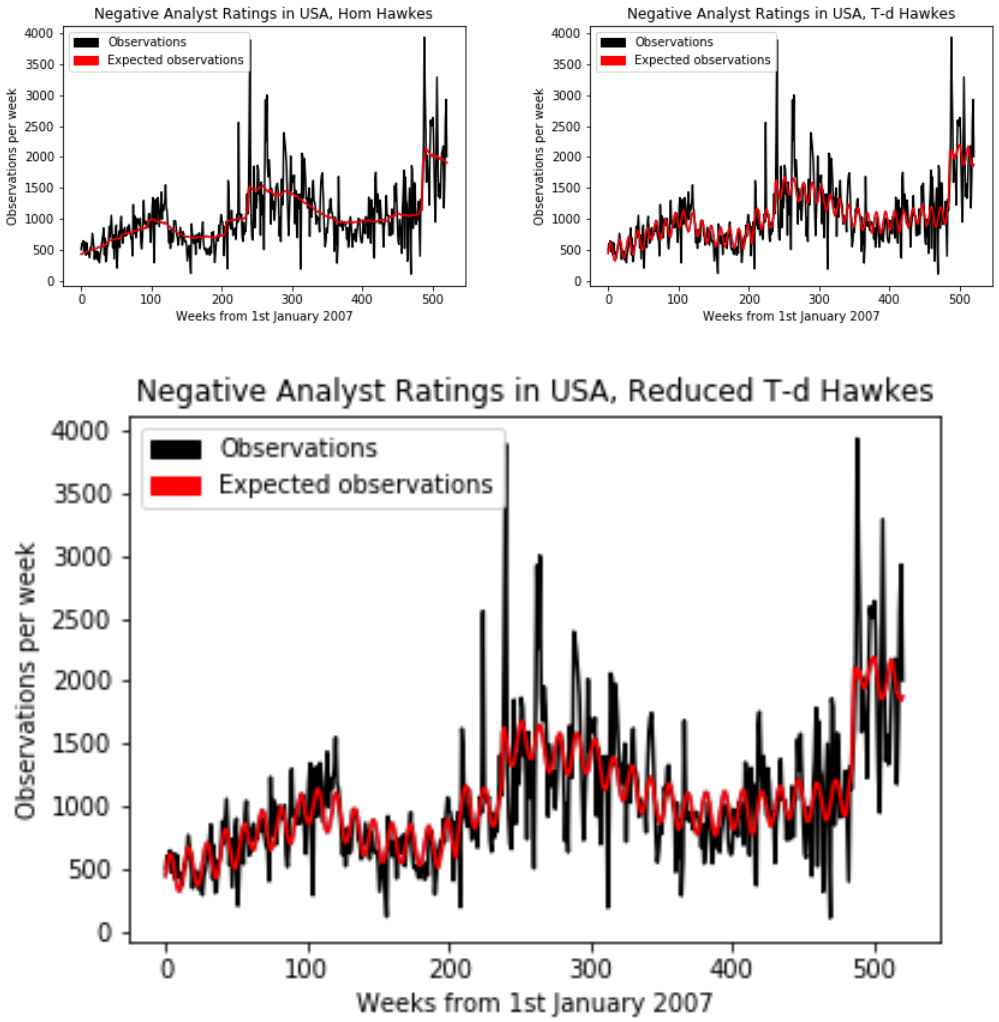


Figure 7.9: The group Negative Analyst Ratings plotted against the expected number of articles of the different models.

At week 240 there is also a big excitation. This time from Negative Security in the Nordic. These are both groups that got through our cleaning of the data since they have observations often but a few times extreme excitations. It is a difficult question if these should be in the results or not. To get a better understanding for this we will examine the excitations a bit more in detail for this complex model than for the smaller models.

7.3.1 Excitation Between Classes

When splitting up the topics and regions the amount of groups though are too many to show in a single graph and get an interpretable result. Therefore only the most important excitations are shown in figure 7.10. In the picture the colours represent a geographical area that are slightly bigger than the 34 regions to see if regions close to each other are more exciting inside the big region.

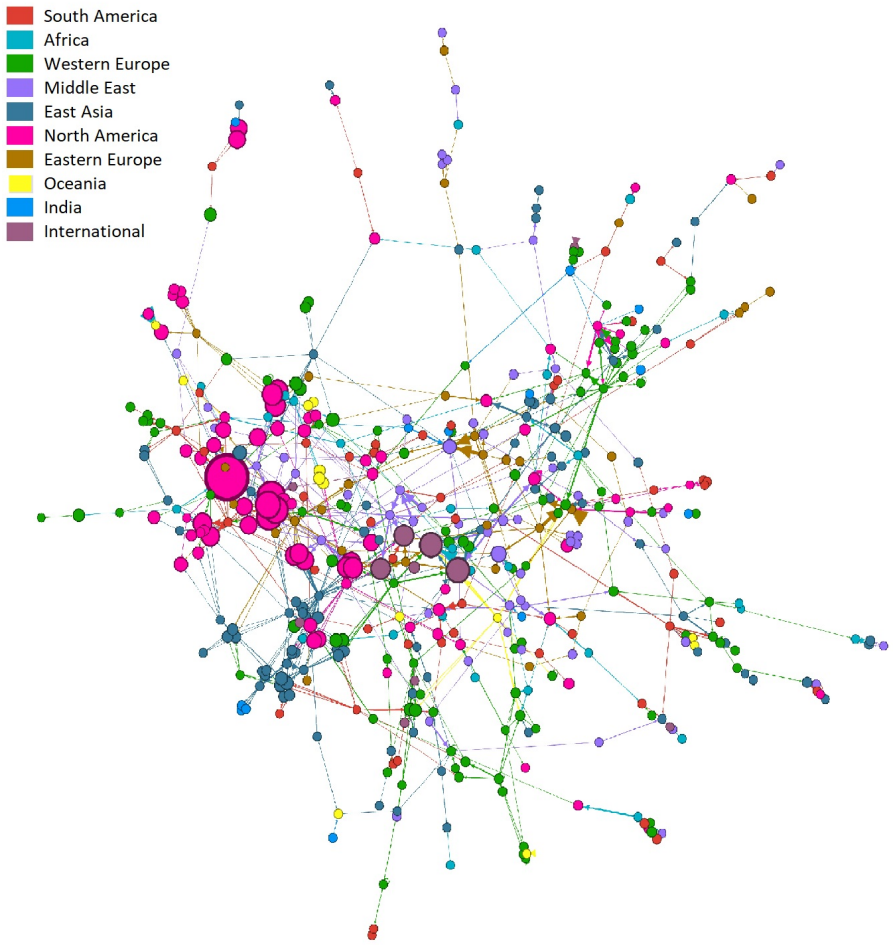


Figure 7.10: A visualisation of the dependencies between the news groups for the 878 most important excitations in the Reduced Time-dependent Hawkes model. Each big region is represented by a unique colour. The size of the group is represented by the size of the node and the size of the excitation by the size of the arrow.

As in the other node maps it is produced using the method Force Atlas 2. This means that nodes that are affected by the same things will be together. In the

figure there is a pink cluster extremely visible and also just under that there is a cluster of East Asia groups. This means that, as hypothesised, the events that occur closer geographically have a tendency to affect geographically close areas to a certain degree. The green colour of Western Europe is also very clear in the picture but spread out to a larger degree. In the plot only the 878 most impactful excitations are plotted and therefore only 552 nodes out of the 2345 nodes are in the picture. Because of this the regions that are in the picture are the ones that affect the news flow the most and North America, East Asia and Western Europe seems dominant in the news flow. In the node map a few important source nodes can be found. Some of these will be examined closer in the following sections.

7.3.1.1 Brexit

In the top right corner of figure 7.10 there is a green cluster of Western European nodes. The central groups in that cluster are Negative Economic Union in Northern Europe and Negative Economic Union in Great Britain. The Northern European node zoomed in can be seen in figure 7.11 where the region colours are the same as in figure 7.10.

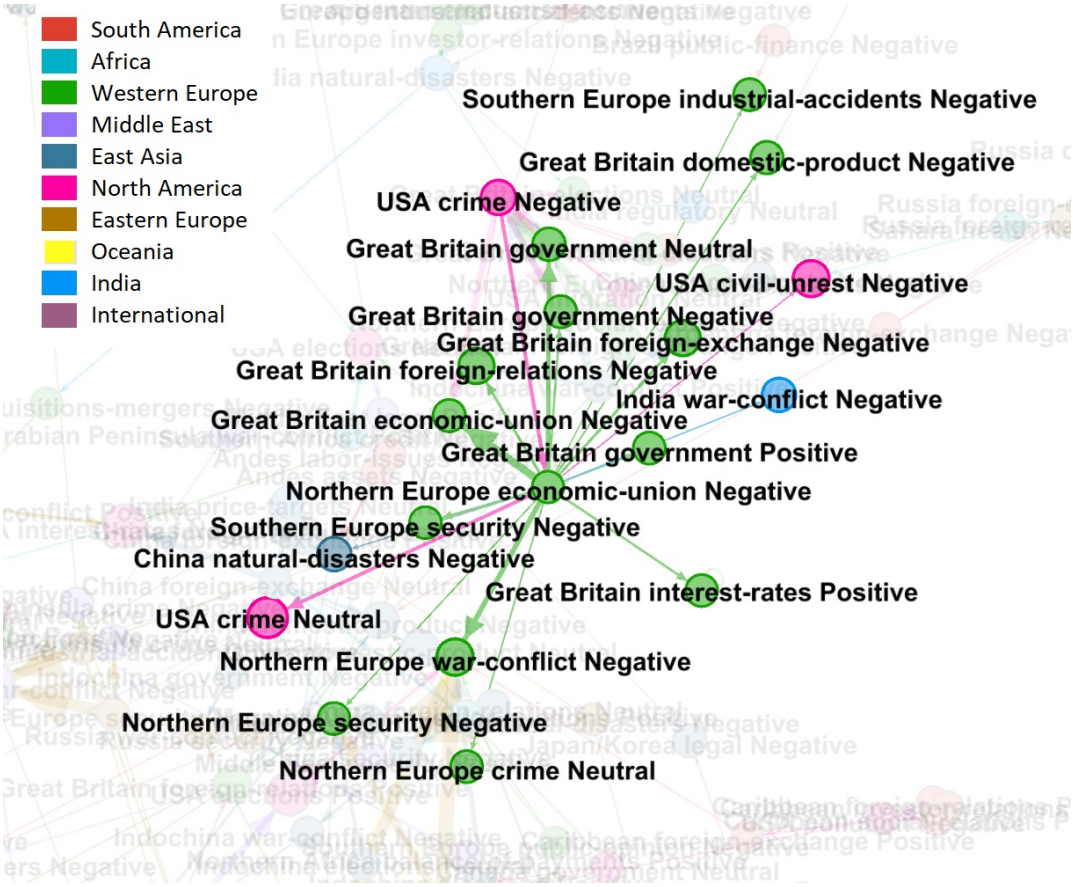


Figure 7.11: A visualisation of the dependencies between the news group Northern Europe Economic Union Negative and its dependent groups. This group has a couple of weeks during Brexit where the news flow explodes.

In the figure it is very obvious that Brexit mostly affected the European news flow based on the colour of the nodes. Secondly one can look at the actual group labels in the figure and see that most topics that are related are economic and government related news. Another thing to note is that most negative news seem to be excited by the Brexit news. This may indicate that experts generally think that Brexit is negative for the markets.

Lastly there seem to be excitation from unrest and crime in USA. This is one of those excitations where one should use caution. The Economic Union news flow is extremely centered around the events following Brexit. This means that if something happened in USA just before the algorithm will find that excitation to be certain. This is one of the hard problems when cleaning the data. If the important extreme events are left out the dynamics of Brexit will never be found but if the

data is kept some weird connections may be suggested by the model. However it is very hard to determine if this is a coincidence or if the Brexit supporters got increasingly eager because of unrest in the world.

7.3.1.2 Arab Spring

Another event that happened one time in the recent history is the Arab Spring. In figure 7.10 the news flow in the Northern part of Africa are very much dominated by the break out and aftermath of the Arab spring that started in the end of 2010. In the colour coding the countries concerned are all in the Middle East group even though they are in the Northern Africa. In the middle of the figure the purple nodes representing this event can be seen. The most prominent node of the purple cluster is Negative Civil Unrest in Northern Africa which excitations can be seen in figure 7.12.

Figure 7.12: A visualisation of the dependencies between the news group Negative Civil Unrest in Northern Africa and its dependent groups. This group has a couple of weeks during the Arab Spring where the news flow explodes.

The node map shows that the civil unrest and the government related news are very related. Considering that much of the events revolved around revolutions this is to be expected. However in the map it does not show any implications of the outer world. Since only about 1 percent of the excitations in even the reduced model is shown in the figures there might be less significant relationships not visible. Moreover the Hawkes model will only use the optimal excitations. In figure 7.13 the excitations of the node Negative Government news in Northern Africa is shown instead.

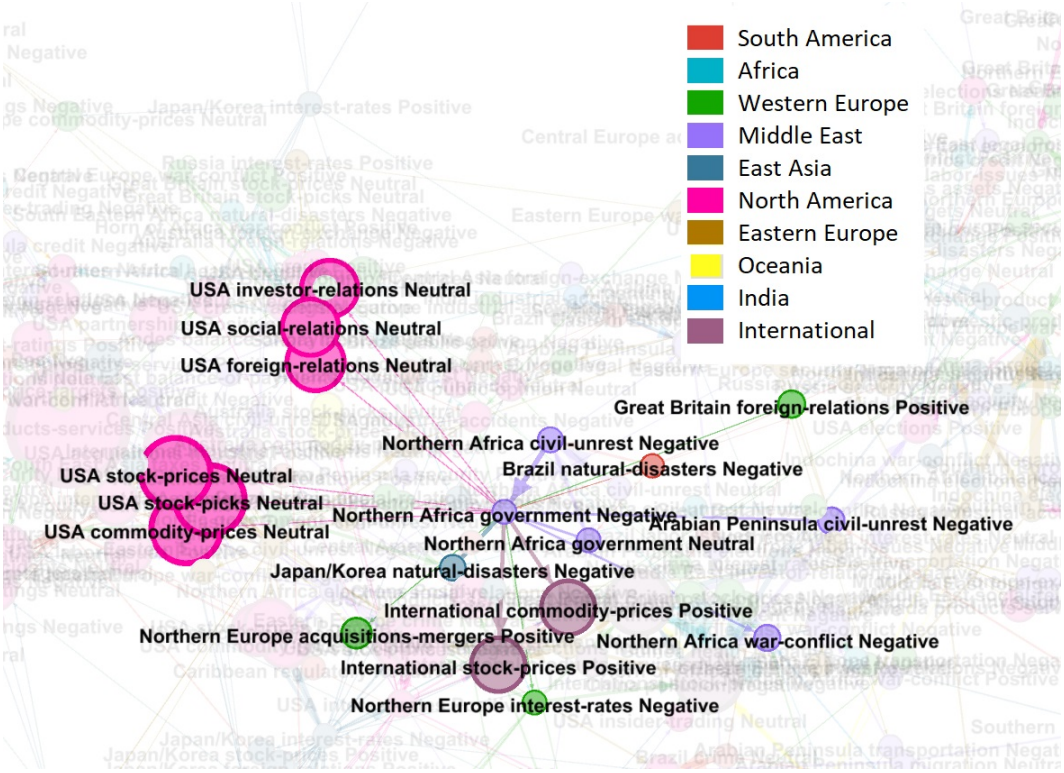


Figure 7.13: A visualisation of the dependencies between the news group Negative Government in Northern Africa and its dependent groups. This group has a couple of weeks during the Arab Spring where the news flow explodes.

This group is also extremely focused around the Arab Spring time line. It seems that most of the excitations to the outer world was put on this group instead though, even though it is the same events that were the origin of the excitation. This is a very important thing to note when examining the results. If there are many groups that will pick up the same events the excitation some groups will not have the expected excitation because another group takes care of it.

Looking at the result of this node instead the interesting interaction that negative government news seem to affect some commodity and stock prices positively. Since the arrows are quite large in comparison to the size of the brow nodes that excitation is also quite a big part of the total news flow in those groups. That can indicate that the Arab Spring had a huge impact on several markets. It can be added that another African one time event, the ebola epidemic is another big influencer of the groups International commodity prices and stock prices. However in that case the connection is even larger and it is affecting the negative counterparts of the groups.

7.3.1.3 East Asia stocks

These last examples are very important for the reader to grasp how the model captures some known behaviour in the news flow. This is probably the most interesting fact about the results for both Brexit and the Arab spring. Knowing this is nice since the amount of variables in the model makes it hard to say anything concrete using confidence intervals or statistical significance of the parameters. Instead we now know that some happenings in the news flow are captured in an intuitive way and therefore the unintuitive but strong excitations may be more trusted. These unintuitive, or not known beforehand, excitations are what makes this study interesting in the first place. One could for example use the information of the news flow to predict how a crime wave in a certain area would affect other regions if there are any excitations pointing towards that. Another way to use the information would be for investors to see how different markets excite each other to understand when to invest. With the help of figure 7.14 one such investment strategy could maybe be formed.

Figure 7.14: A visualisation of the dependencies between the news group Positive Stock Picks in Indochina and its dependent groups.

It seems that when the stock prices goes up in the region Indochina, experts will pick stocks there as recommendations. Then in turn the stock and commodity prices in China and India goes up. This excitation is very strong in relation to how small the groups are as is suggested by the big arrows and small nodes. Since the decay

times are in weeks this information could maybe be used in some smart way where if the Indochina stock market is thriving we buy chinese and Indian stocks and hold a few weeks. If the information can be used to formulate a good trading strategy will not be investigated further in this thesis though but could be an interesting topic to pursue.

7.3.1.4 Interest rates

In the world of macro economics the interest rates play a big role. This can be seen in the model since many of the big emitters are interest rate reports from the big regions. In figure 7.15 the positive sentiment news about the Interest Rates in USA is shown.

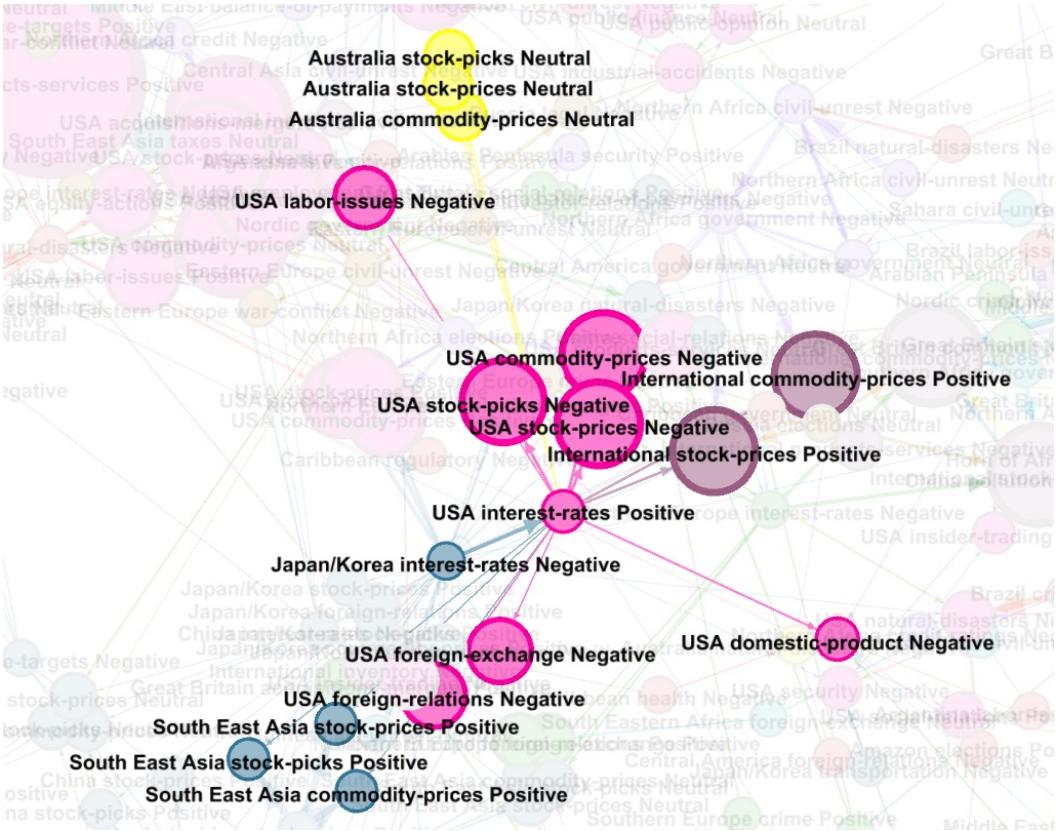


Figure 7.15: A visualisation of the dependencies between the news group Positive Interest Rates in USA and its dependent groups.

Various groups seems to be affected by it, mostly in USA and Asia. It is in turn excited by the negative sentiments of the Japanese Interest Rates. That node can be seen in figure 7.16.

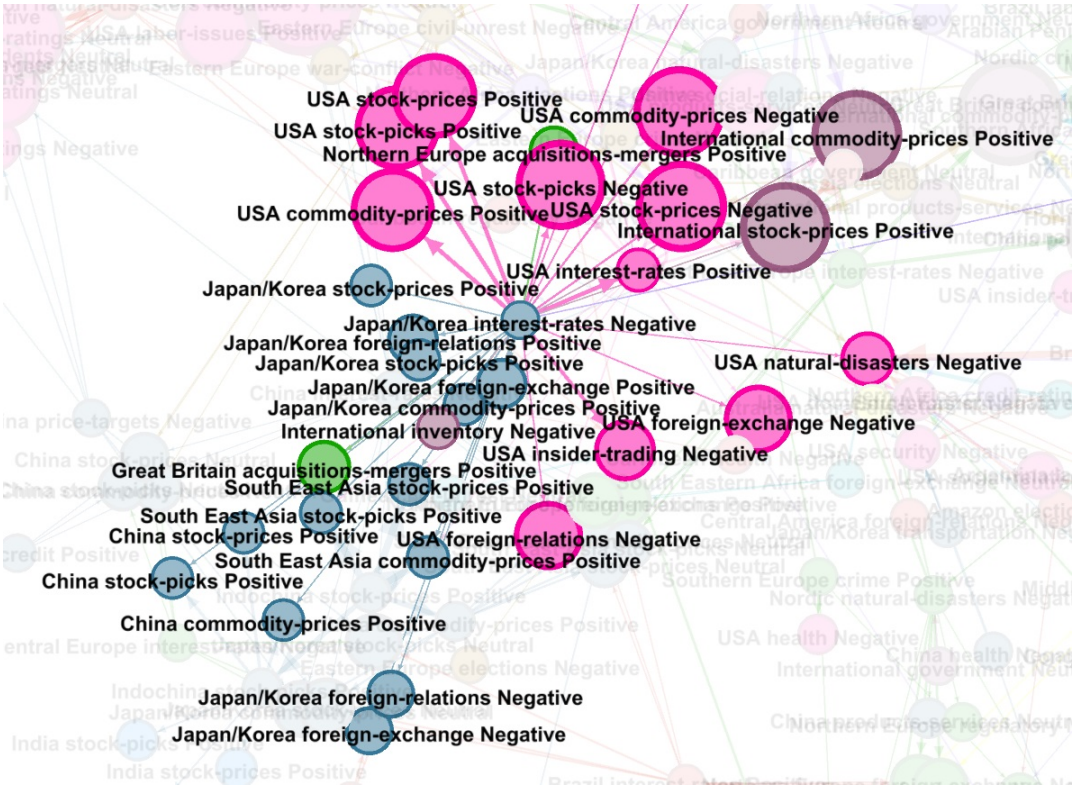


Figure 7.16: A visualisation of the dependencies between the news group Positive Interest Rates in Japan/Korea and its dependent groups.

Even though both groups are relatively small they excite a lot of economic news in primarily America and Asia. They both also have an effect on the International stock and commodity market. Looking into that more in detail in figure 7.17.

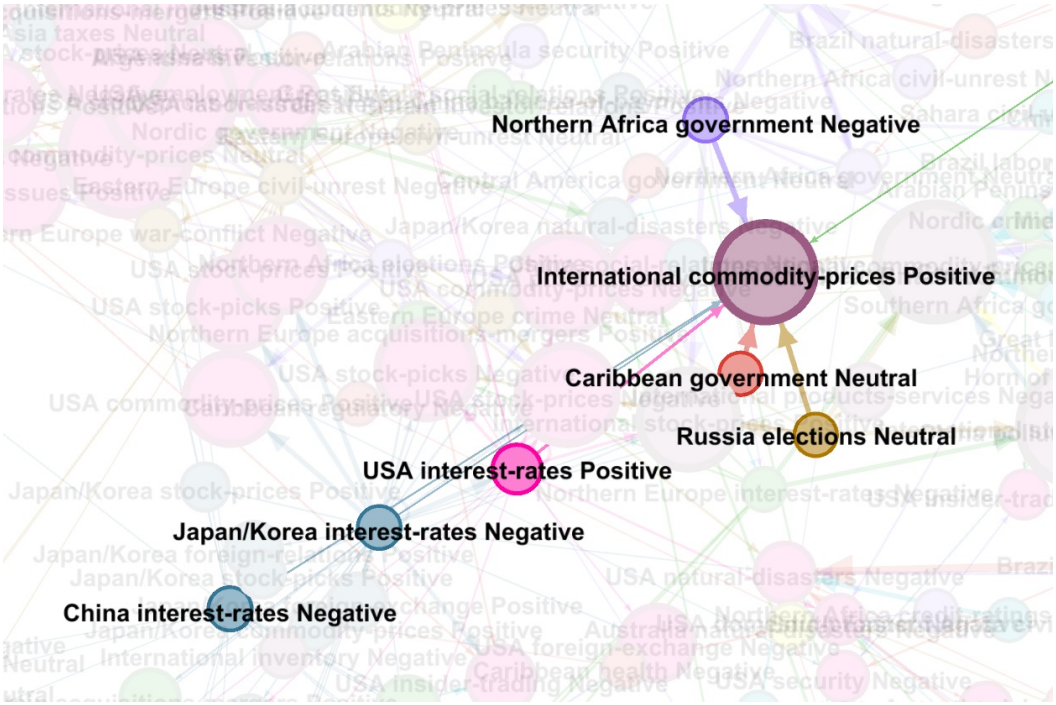


Figure 7.17: A visualisation of the dependencies between the news group Positive Commodity Prices in International group and its dependent groups.

One can see what was noted before that the Arab Spring had an effect on this market as well. There also seem to be some government related groups in Caribbean and Russia and finally the Interest Rates of China are also important. It seems that the interest rates in general has a small effect on many things but that more specific events can affect the flow more than every small interest rate report. This is also reasonable since there are only a few reports that actually should chock the market and that most reports of the interest have little to no effect.

7.3.1.5 Conflicts

The news flow are of course not only economic reports of different kinds. Some things related to the conflicts in the world can also be found, more general than the effects of the Arab Spring. If moving in the opposite directions of the arrows there will eventually be a group without excitation towards it, that maybe affects many groups. In the economic news that was the interest rates. In the conflicts department the group Transportation in Russia is one of those nodes as well. It can be seen in figure 7.18.

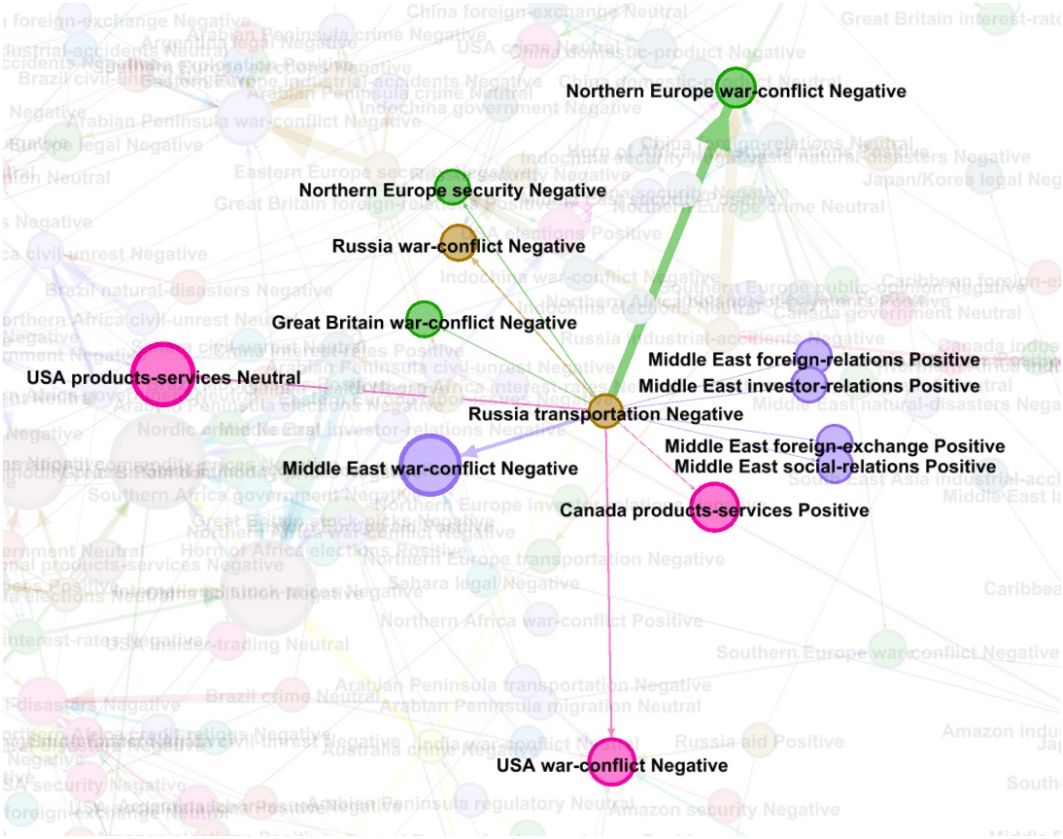


Figure 7.18: A visualisation of the dependencies between the news group Negative Transportation in Russia and its dependent groups.

Russia has been relatively close geographically to many recent conflicts, that took place in the Middle East and they are just as USA a big power that always are mentioned in news in times of uncertainties. It is therefore not surprising that they have a big presence in the department. The excitation towards Northern Europe may seem unexpected. An explanation may be that even though Northern Europe are not often mentioned in conflict related news the closeness to Russia makes every report from Russia very important even for Northern Europe.

We will end this whole node map examination with a last node map of the conflict in the Middle East. In the figure 7.19 the excitation from Transportation in Russia can be seen.

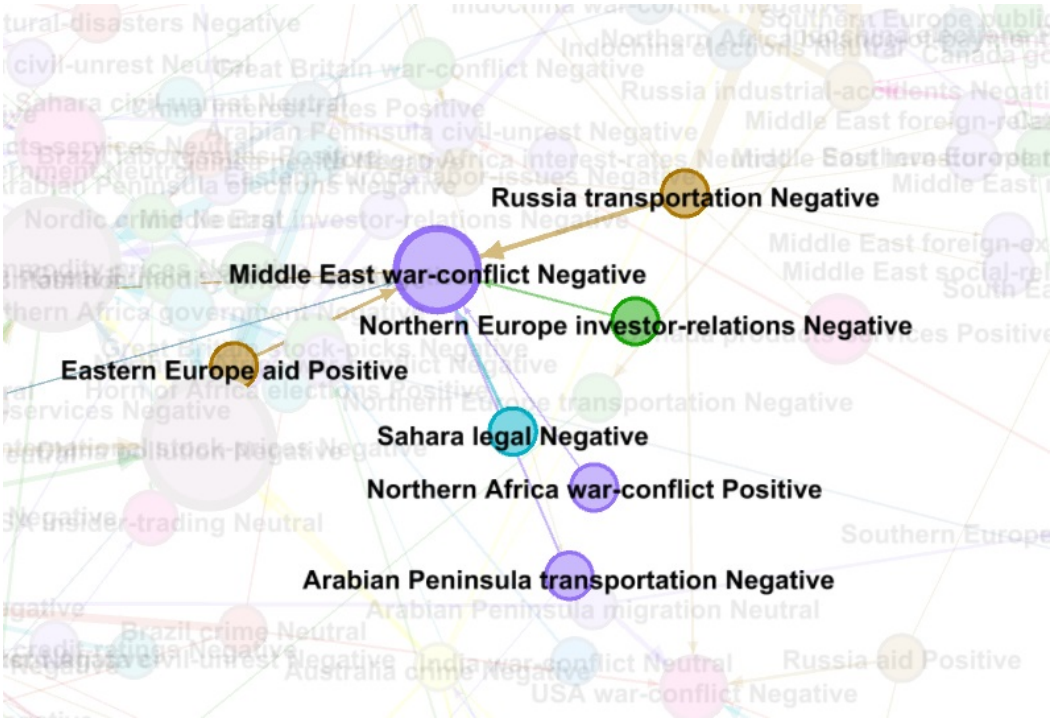


Figure 7.19: A visualisation of the dependencies between the news group Negative War and Conflict in the Middle East and its dependent groups.

It can also be seen that aid from Eastern Europe is a good indicator that something is happening in the Middle East. Also some interaction seem to exist with Northern Africa and the Sahara region. All regions seem to be in the same area as the conflict. As an introduction to the next section it can be noted that the size of the Middle East War group is much larger than the other groups in the figure.

7.3.2 Size of Groups

Lastly we need to touch upon one detail that has been emerging in the results. In the chapters describing how the data was cleaned it was mentioned that one of the side effects of taking away the most periodic groups, the largest groups were also taken away. This was good since the groups became more balanced when taking away both the smallest and biggest groups. The attentive reader has seen in most of the node maps shown though that still the small groups excite the big groups and almost never the opposite way around. This is not something that is preferable since we lose the nice causality property of the model if there can only be excitation one way between two groups.

In the last figure concerning the Middle East conflicts it may be more sensible that

the war excite the aid and not the other way around. The really big exciters, the interest rates, were all small groups that often excited larger groups. There are some examples of almost equally sized groups having excitation to each other but never when the groups are of different size. This is another thing that has to be taken into consideration when watching these node maps. If the nodes are not of equal size the direction of the arrow can only be in one direction it seems. Fortunately in the East Asia stock example where we want to use the direction of the arrow the groups are of equal size so this complication should not affect that example.

7.4 Overlapping Classes

Some results above indicates that the amount of groups are too many in the tests above, for example in figure 7.10 where it seems that the continents excite themselves in a large part. One could think that in the news groups many of the economic news may report the same event many times why they should also behave the same way. The problem here is how to change the classes without interfering too much with intuition so the result gets tampered with.

For this we have the other more complex model where all news exist in a space and the groups are densities in this space rather than categorical variables. The results for this model when the news topics and regions are made up of only 5 densities each can be seen in table 7.4. Only the Homogeneous Hawkes model was tested with in the Overlapping Classes framework. This is because the arguments that will be made further down will apply for all models but are easiest displayed with that model.

Table 7.4: Likelihoods of different models. There are 171 728 259 observations.

Model	Log Likelihood	Parameters	BIC
Homogeneous Discrete Hawkes	-606 719 122	34104	483 078 266

The results indicates that likelihood favour the more advanced Overlapping model when considering that there are very few classes. The reason why the amount of classes is small is that when all classes are used this model becomes identical to the distinct model almost. The reason to use this model in the first place is to reduce the amount of classes. One problem here is that the new classes became very uninteresting. Instead of clustering together groups with similar behaviour it seems to be clustering together groups so that the new groups are easy to predict the intensity of.

This problem has its origin in that the estimation of the intensities for the more feisty groups are not that good, so even if the spatial penalty will be big it is still favourable to just create easy groups with an almost constant news flow. The easy

estimates will then prove to not fit the original classes at all when using the new projected class on the original group space, see figure 7.20.

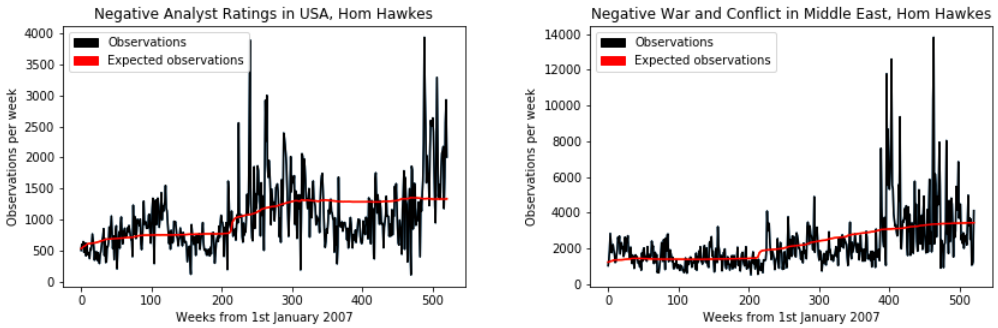


Figure 7.20: To the left: The group Negative Analyst Ratings in USA plotted against the expected number of articles of the Homogeneous Hawkes model. To the right: The group Negative War and Conflict in the Middle East.

The problem is that the intensity is very close to a constant intensity even though the group intensity is changing. This is because the model will cluster together and then say how much of the news that should be in every group but only with a scaled intensity. This makes for a relatively high likelihood but an uninteresting result. We can in the figure see that both War in the Middle East and Negative Analyst Ratings in USA have much the same shape of the intensity. This is because the get most expected observations from the same class density. This would be good if their observations had the same shape. This seems to not be the case however which means the model has clustered together groups that are not behaving the same. If the model were more exact this would not happen because then the spatial penalty of clustering together groups with different behaviour would be too great. There was one class that caught the Arab Spring though, see figure 7.21.

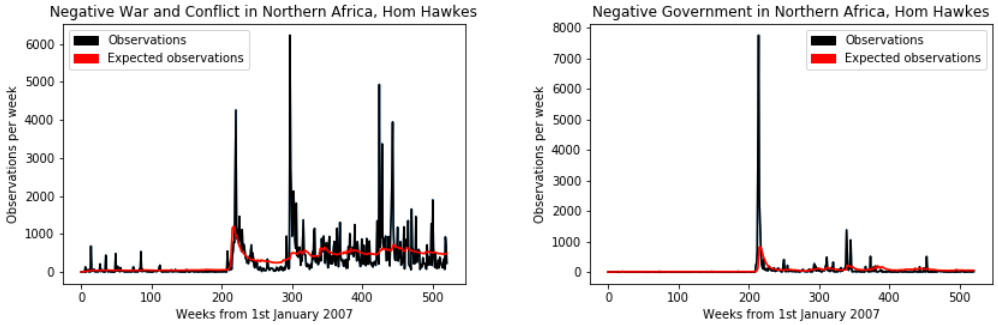


Figure 7.21: To the left: The group Negative War and Conflict in Northern Africa plotted against the expected number of articles of the Homogeneous Hawkes model. To the right: The group Negative Government in Northern Africa.

In this figure we see two war related groups in Northern Africa that has been clustered together. This group got all the groups with a clear Arab Spring peak which indicates that for obvious enough similarities the model works. Therefore it is plausible that the extensions that will make this model a contender to the distinct model are not far away. It is also possible that the amount of classes in this overlapping test were too few to get interesting results. This may call for additional tests that were not conducted in this thesis.

This result therefore has to be seen as an extension to the distinct class result in the sense that the classes there may not be optimal since the likelihood in the overlapping case is indeed better when considering the amount of classes used. Though for the results produced in this thesis the distinct class case has better interpretability in terms of excitations and is for all intents and purposes likely a better model. However, in this thesis the ground work of the model in terms of algorithms and model can be seen as a result itself. If a more exact fit is found for a distinct model, the extensions to make it overlapping are probably the same and may prove to give better results.

7.5 Prediction

As a last note in the result department the predictive power of the Distinct Classes model will be discussed. This will not be an extensive quantitative result. This section is supposed to give the reader an understanding just how bad the overfitting problem is for the models. In this thesis the aim is to see the connections. This section will reach just outside the scope and see what is beyond.

In the figure 7.22 the same result as in 7.9 is shown with the additions of expected observations when training until the end of 2012, 2014 and 2015 respectively.

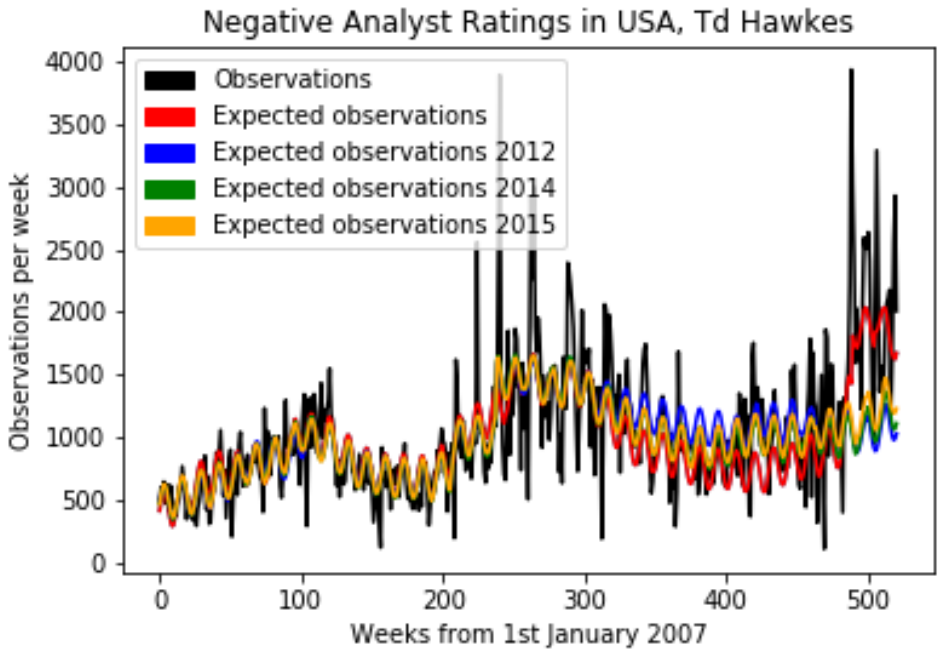


Figure 7.22: The observations of the group Negative Analyst Ratings in USA plotted against the expected number of articles the Time-dependent Hawkes model with 2345 groups. The different colours represent in the end of what year the training data stops at. The rest of the data is unknown for the model.

In the figure all intensities are following each other in the beginning but the blue curve of 2012 is really guessing too high around week 390. They all converge again at around week 480 but then around week 490 Brexit hits. Because the red curve is trained on the whole data set it can capture it. All the other three are missing it completely. The later the training stopped the higher the guess is after week 500 though so it is possible that some hints of Brexit were caught in the 2015 model that were not caught in the 2012 model. The differences are extremely small though. One obvious example of that the model may be overtrained for predictive purposes can be seen in figure 7.23.

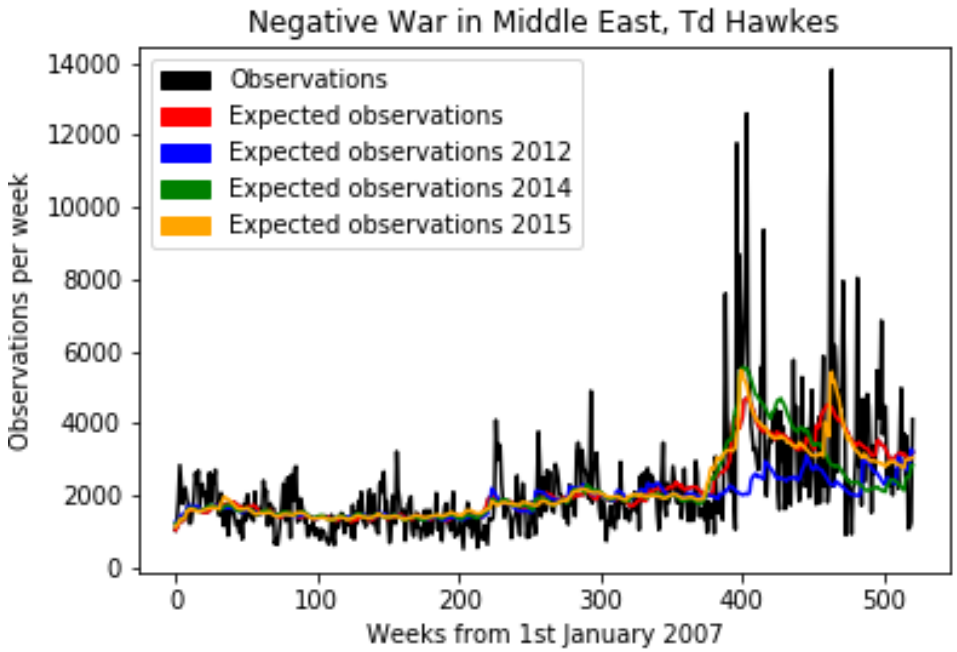


Figure 7.23: The observations of the group Negative War and Conflict in the Middle East plotted against the expected number of articles the Time-dependent Hawkes model with 2345 groups. The different colours represent in the end of what year the training data stops at. The rest of the data is unknown for the model.

In the figure the War and Conflict in the Middle East is once again examined. There are 2 big peaks, one around week 400 and one around week 470. This is autumn and winter 2014 and 2015 respectively. This means that the blue curve, trained until end of 2012 will not have any of these peaks in the training. As can be seen it misses both peaks. The green curve, trained until end of 2014 will have the first peak in its training set. The second peak is completely missed by it. For the last peak only the red and yellow curve picks it up and those were trained on sets containing it. This is extremely bad if the model were to be used as a predictive model.

To get a comparison the different test sets are plotted for 2 economic groups in the simpler model with all countries in the same bucket as well. This can be seen in figure 7.24.

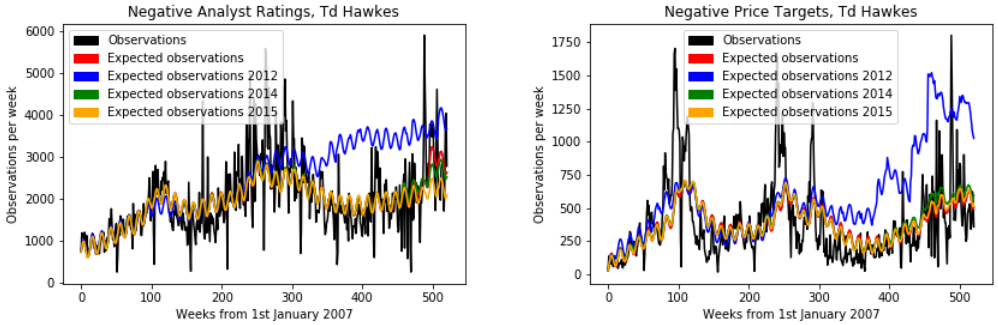


Figure 7.24: The observations of the group Negative Negative Analyst Ratings plotted against the expected number of articles the Time-dependent Hawkes model with 53 news groups. The different colours represent in the end of what year the training data stops at. The rest of the data is unknown for the model.

Here we see that the 2012 model is really bad on the test data. However in contrast to the model with countries it actually happens things after the training data set stopped. In the previous case almost all excitation were in the training part of the data set. This points towards that our cleaning of the data for that model should have been even stricter. Groups with an extremely non-stationary distribution should maybe have been left out.

Because all countries are placed in the same bucket, this phenomenon is less apparent with the simpler model, meaning that this problem is less visible. Of course with the addition that complex connections cannot be found at all. This small quality section about prediction will conclude the result section and hopefully the reader realises that depending on what the use of the model is, a more complex or simple model could be favoured.

Chapter 8

Discussion

In this chapter there will be a brief discussion of the project in general and the results in chapter 7. Also some outlook and possible further work will be proposed.

8.1 Results

8.1.1 Numerical Optimisation

Most of the work has been related to estimate the parameters using the gradient methods to maximise the likelihood. The computational problems because of the massive amount of data and groups led to some of the more important decisions of the project. The final 2345 groups, after the data set was cleaned, could not have been many more with the same hardware even if the data set had been larger. With the single GPU that were used every likelihood function took about 5-10 seconds to calculate and convergence were not reached until at least 20-50 thousand iterations which is a couple of days. With more groups the results could take weeks to produce.

To reduce the amount of time every single model took to train, transfer learning were applied after the models were trained to the applicable parameters. The excitations and base intensities found in one model were used as starting values for the other more computationally hard models. As an extension this could possibly have been used to speed up computations even further if transfer learning were to be applied from training on a shorter time span. One such algorithm could training on 1 month, then 1 year and then 10 years. Convergence should be much faster this way and it would enable the possibility of a more complex model if there should be an application for it. As it is now though we think that more data would be needed if the data is split up much more to avoid the problems with few observations touched upon earlier.

Closely linked to the subject of transfer parameters from shorter learning data set

is the predictive power of the model. Some important observations that were not in the shorter time span will not get trained on. This could have been researched more in this thesis. How the different training sets affect the end result. It is obvious that the Brexit events will not be caught in the model if the years of the event is not in the training data. This is of course a big problem with the model. There will always be events in the future that the model can't predict. It would have been interesting to measure the test accuracy versus the training accuracy more quantitatively than what was done.

One last note about the estimation of the parameters is that it is peculiar that the reduced model have a better likelihood than the full model in all optimisations. This is of course an impossibility if the optimisation went perfect since the reduced model is a subgroup of the full model. This might be because that it is hard for the optimiser to estimate the parameter to exactly 0 and because of the number of parameters even a small difference from 0 tampers with the result. A solution to it might be to lower the learning rate for the full model. As it is done in the project the same learning rate is used in all estimations.

8.2 Interpretation of Results

When dealing with as many parameters as has been done in this project the significance of the estimated parameters is a big concern. How to distinguish a weird coincidence from an extremely important single event is almost impossible. This has been a struggle from the beginning. By only accepting large enough excitation the numerical errors will be few but weird random correlations may still exist. This is something the reader has to take into consideration when examining the results.

We really do think that a large amount of the information can be trusted though. This is because there are an abundance of intuitive results. This is an indication of that the model works and we will therefore probably be able to use the unexpected results, at least if they are strong.

As for what model we would suggest, the reduced model seems like a very nice option. There is a possibility to tune it to the size of excitations that are wanted and it seems that Hawkes in general actually gives a way better fit than the Poisson case. This is also an indication that the hypothesis works, that the Hawkes model is a model that can be used for the news flow in the first place.

One thing to remember when examining the results is that the decay time of the excitation kernel is pretty important to get sensible results. If the decay time tend to infinity the resulting intensity tend to a step function. This is the reason that the inhomogeneous model added a trend, but since not all growing trends are linear

this problem is a recurring one. To deter this behaviour further the half-time for the decay was capped at 2 years. This was set pretty high to not interfere with the optimisation too much but it seems that a few groups reached that cap anyway. This may be an indication of that there are extremely long correlations or that it is just approximating a step function. These results should be used with caution in our opinion. Most groups had a half-time of 2-10 weeks and these results should probably be trusted more in general. This is a reason why in the node graph the weights are not scaled with decay. Most sensible decay time are almost the same and some weird long interactions may have gotten a huge weight in the node graph. There is of course the option to further limit the half-time but the trouble is that we do not know if some news have longer dependence. There is a possibility that an election could have effects for 4 full years rather than 5 weeks as an example.

8.3 Outlook and Further Work

During this thesis there were a lot of things that were ultimately dropped because of time constraints or that they would widen the scope of the project too much. There were also many paths that were not pursued since the thesis in mathematical statistics and not in finance or social studies. However these subjects would be very interesting to study based on the results in this thesis.

8.3.1 Extending the Model

There are many ways in which one could make our model more complex and sophisticated by only simple modifications. By changing the time dependent background intensity of the Poisson Process or changing the excitation kernel the whole model could change drastically. It would be interesting how such a model would compare to our model.

For the overlapping model in our thesis only the surface were scratched in this thesis. Our tests indicate that the overlapping model may be redundant since the distinct model will exhibit more interesting excitations. With a more exact excitation kernel this may be avoided and this method may be useful.

For both models it would be very easy to check whether the excitation and correlation between classes give the same results. Otherwise it may be possible to cluster together topics not by hand or with likelihood optimisation but with how much correlation there are between them. The thought process behind this is that if they are highly correlated they may be very much alike and therefore their excitations should be the same. This also gives birth to another interesting subject which

is if our findings could have been found using simple correlation or if the causality that the Hawkes Process provides actually is significant.

An even more complex model would be to have a completely different approach to the problem and use the news flow as observations of a hidden process that is the real world. This is probably more close to the truth about how the news flow works but the computational problems may be even greater and the direct interpretability from the excitations may be lost in such a model.

8.3.2 Other Work

The results from this thesis could very easily be used in other studies. As mentioned in the introduction the step to model the news flow may not be the last step since this information should be used in some way.

In an era where machine learning and neural networks are very prominent it is very tempting to plug in the intensity curve of a few topics and regions and see if just the general news flow can predict things. Since this thesis is done at a hedge fund it is close at hand to try this out on different markets. An idea is to identify which news that affect the negative and positive economic news and use them as features in a neural network. Another idea is to directly look at the markets. As stated in the results the Asian markets seem to be very related according to our results. Since there are causality in the model it should be possible to establish a trading strategy based on this information. A side note to this though is that the algorithm could probably have been used on the markets directly if there is such a relationship.

For a completely different sort of continuation we think that a social study based on the results could be interesting. Increasing unrest in the world and migration streams could possibly be studied based on how the war haunted regions excite the rest of the world. Also political and government related excitations could carry some indicative power in how the people perceive the current situation of the region.

Chapter 9

Conclusions

In this thesis our primary goal was to investigate how different parts of the news flow affected each other. To see if there were any patterns in the data a few different methods were compared. As a base case the Homogeneous Poisson Point Process was chosen.

Adding excitation to the Homogeneous Poisson Point Process makes it a standard form Hawkes Point Process. This model did generally increase the performance by a big margin and would be possible to use. However a problem that were recurring was that much of the information that were captured with the model is possibly redundant since it used excitations to model seasonality and the growing trend in the data. For this thesis this was misleading since the excitations hinted at close relations between topics and regions just because their news flow is increasing in time.

To avoid this behaviour an Inhomogeneous Poisson Point Process was tried out as well. The time dependence in it consisted of a growing trend and a seasonality of a quarter of a year. When adding excitation to this process it yielded a result that were more suited for our purposes.

Finally to add as much predictive power as possible to the complex Inhomogeneous Hawkes Process some excitations that were very small were deemed exactly 0 to decrease the amount of variables that were estimated. By doing this the complex model was always favoured in terms of likelihood and by the Bayesian Information Criteria in front of the simpler models.

With the Inhomogeneous Hawkes Process node maps with the most important relations could be constructed. The result seem to capture known events in the recent history such as Brexit and the Arab Spring. Also there seems to be geographical dependence in the sense that more countries in the same general area has an affect

on each other compared to countries far away.

Looking specifically on the topics, the economic news seem to be closely related to each other with very big excitations between some of the topics. Other topics that had a big connection were for example natural disasters and migration. Government related areas such as foreign relations and public opinion also were clustered together by the algorithm.

Chapter 10

Appendix A

Table 10.1: Country codes Part 1

Code	Country	Code	Country	Code	Country
AD	andorra	CF	central african republic	GD	grenada
AE	united arab emirates	CG	congo	GE	georgia
AF	afghanistan	CH	switzerland	GF	french guiana
AG	antigua & barbuda	CI	ivory coast	GG	guernsey
AI	anguilla	CK	cook islands	GH	ghana
AL	albania	CL	chile	GI	gibraltar
AM	armenia	CM	cameroon	GL	greenland
AN	netherlands antilles	CN	china	GM	gambia
AO	angola	CO	colombia	GN	guinea
AQ	antarctica	CR	costa rica	GP	guadeloupe
AR	argentina	CS	serbia and montenegro	GQ	equatorial guinea
AS	american samoa	CU	cuba	GR	greece
AT	austria	CV	cape verde	GS	south sandwich isl.
AU	australia	CW	curacao	GT	guatemala
AW	aruba	CX	christmas island	GU	guam
AX	åland	CY	cyprus	GW	guinea bissau
AZ	azerbaijan	CZ	czech rep.	GY	guyana
BA	bosnia and herzegovina	DE	germany	HK	hong kong
BB	barbados	DJ	djibouti	HM	mcdonald islands
BD	bangladesh	DK	denmark	HN	honduras
BE	belgium	DM	dominica	HR	croatia
BF	burkina faso	DO	dominican republic	HT	haiti
BG	bulgaria	DZ	algeria	HU	hungary
BH	bahrain	EC	ecuador	ID	indonesia
BI	burundi	EE	estonia	IE	ireland
BJ	benin	EG	egypt	IL	israel
BL	saint barthelemy	EH	western sahara	IM	isle of man
BM	bermuda	ER	eritrea	IN	india
BN	brunei	ES	spain	IO	br. indian ocean
BO	bolivia	ET	ethipoia	IQ	iraq
BR	brazil	EU	european union	IR	iran
BS	bahamas	EZ	euro zone	IS	iceland
BT	bhutan	FI	finland	IT	italy
BV	bouvet island	FJ	fiji	JE	jersey
BW	botswana	FK	falkland islands	JM	jamaica
BY	belarus	FM	micronesia	JO	jordan
BZ	belize	FO	faroe islands	JP	japan
CA	canada	FR	france	KE	kenya
CC	cocos islands	GA	gabon	KG	kyrgyzstan
CD	dem. Republic of kongo	GB	great britain	KH	cambodia

Table 10.2: Country codes Part 2

Code	Country	Code	Country	Code	Country
KI	kiribati	MY	malaysia	SG	singapore
KM	comoros	MZ	mozambique	SH	saint helena
KN	nevis anguilla	NA	namibia	SI	slovenia
KP	north korea	NC	new caledonia	SJ	svalbard
KR	south korea	NE	niger	SK	slovakia
KW	kuwait	NF	norfolk islands	SL	sierra leone
KY	cayman islands	NG	nigeria	SM	san marino
KZ	kazakhstan	NI	nicaragua	SN	senegal
LA	laos	NL	netherlands	SO	somalia
LB	lebanon	NO	norway	SR	suriname
LC	saint lucia	NP	nepal	SS	south sudan
LI	liechtenstein	NR	nauru	ST	sao tome
LK	sri lanka	NU	niue	SV	el salvador
LR	liberia	NZ	new zealand	SX	sint marteen
LS	lesotho	OM	oman	SY	syria
LT	lithuania	PA	panama	SZ	swaziland
LU	luxembourg	PE	peru	TC	caicos islands
LV	latvia	PF	polynesia	TD	chad
LY	libya	PG	papua new guinea	TF	french s. territories
MA	morocco	PH	philippines	TG	togo
MC	monaco	PK	pakistasn	TH	thailand
MD	moldova	PL	poland	TJ	tajikistan
ME	montenegro	PM	st. pierre and mi.	TK	tokelau
MF	saint martin	PN	pitcairn island	TL	east timor
MG	madagascar	PR	puerto rico	TM	turkmenistan
MH	marshall islands	PS	palestine	TN	tunisia
MK	macedonia	PT	portugal	TO	tonga
ML	mali	PW	palau	TR	turkey
MM	myanmar	PY	paraguay	TT	trinidad and tobago
MN	mongolia	QA	qatar	TV	tuvalu
MO	macau	RE	reunion	TW	taiwan
MP	n. mariana islands	RO	romania	TZ	tanzania
MQ	martinique	RS	serbia	UA	ukraine
MR	mauritania	RU	russia	UG	uganda
MS	montserrat	RW	rwanda	UM	USA minor isl.
MT	malta	SA	saudia arabia	US	USA
MU	mauritius	SB	solomon islands	UY	uruguay
MV	maldives	SC	seychelles	UZ	uzbekistan
MW	malawi	SD	sudan	VA	vatican
MX	mexico	SE	sweden	VC	grenadines

Table 10.3: Country codes Part 3

Code	Country
VE	venezuela
VG	virgin islands (UK)
VI	virgin islands (USA)
VN	vietnam
VU	vanuatu
WF	futuna islands
WS	samoa
XK	kosovo
XX	international
YE	yemen
YT	mayotte
ZA	south africa
ZM	zambia
ZW	zimbabwe

Table 10.4: Regions Part 1

Caribbean	Polynesia	Guinea	Mediterranean	Eastern Europe	Mashriq
AG	AS	BF	AD	BG	AE
AI	CK	BJ	AL	BY	BH
AN	FJ	CI	BA	CS	IL
AW	FM	CM	CY	EE	JO
BB	GS	CV	ES	LT	KW
BL	GU	GA	GI	LV	LB
BM	KI	GH	GR	MD	OM
BS	MH	GM	HR	ME	PS
CU	MP	GN	IT	PL	QA
CW	NC	GQ	MK	RO	SA
DM	NF	GW	MT	RS	SY
DO	NR	LR	PT	UA	YE
GD	NU	MG	SI	XK	
GF	PF	SL	SM		
GP	SB	SN	TR		
GY	TK	ST	VA		
HT	TO	TG			
JM	TV				
KN	VU				
KY	WF				
LC	WS				
MF					
MQ					
MS					
PN					
SR					
SX					
TC					
TT					
UM					
VC					
VG					
VI					

Table 10.5: Regions Part 2

Swahili Coast	Central America	Sahara	Nordic	France Germany
BI	BZ	CF	AX	BE
KM	CC	EH	DK	DE
MU	CR	ML	FI	EU
MW	GT	MR	FO	EZ
MZ	HN	NE	GL	FR
RE	MX	NG	IS	LU
RW	NI	SD	NO	MC
SC	PA	SS	SE	NL
TZ	PR	TD		
UG	SV			
YT				

Table 10.6: Regions Part 3

South East Asia	Middle East	Southern Africa	India	Central Asia	Central Europe
BN	AF	BW	BD	KG	AT
CX	AM	LS	BT	KZ	CH
ID	AZ	NA	IN	MN	CZ
MY	GE	SZ	IO	TJ	HU
PH	IQ	ZA	LK	TM	LI
PW	IR	ZM	MV	UZ	SK
SG	PK	ZW	NP		
TL					

Table 10.7: Regions Part 4

Arctic-America	Indochina	Horn of Africa	Northern Africa	Great Britain
AQ	KH	DJ	DZ	GB
BV	LA	ER	EG	GG
HM	MM	ET	LY	IE
PM	TH	KE	MA	IM
SJ	VN	SO	TN	JE
TF				

Table 10.8: Regions Part 5

Amazon	Argentina	China	Congo	Japan Korea	Oceania
BO	AR	CN	AO	JP	AU
CO	FK	HK	CD	KP	NZ
EC	PY	MO	CG	KR	PG
VE	UY	TW	SH		

Table 10.9: Regions Part 6

Andes	USA	Brazil	Russia	International	Canada
CL	US	BR	RU	XX	CA
PE					

Bibliography

[sto, 2010] (2010). Stock price prediction using financial news articles. *2010 2nd IEEE International Conference on Information and Financial Engineering, Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on*, page 478.

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akadémiai Kiado.

[Asur and Huberman, 2010] Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA. IEEE Computer Society.

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

[Bray and Paik Schoenberg, 2013] Bray, A. and Paik Schoenberg, F. (2013). Assessment of point process models for earthquake forecasting. *Statistical science*, 28.

[Heston and Sinha, 2017] Heston, S. L. and Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, 73(3):67 – 83.

[Jacomy et al., 2014] Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE*, 9(6):1 – 12.

[Karr, 1986] Karr, A. (1986). *Point Processes and Their Statistical Inference*. Probability : Pure and Applied, a Series of Textbooks and Reference Books, Vol 2. Taylor & Francis.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *Published as a conference paper at ICLR 2015*.

[Laub et al., 2015] Laub, P. J., Taimre, T., and Pollett, P. K. (2015). Hawkes processes.

[Mitra and Mitra, 2011] Mitra, G. and Mitra, L. (2011). *The handbook of news analytics in finance*. Wiley and Sons Ltd. Publication.

[Rizoiu et al., 2017] Rizoiu, M.-A., Lee, Y., Mishra, S., and Xie, L. (2017). A tutorial on hawkes processes for events in social media. *arXiv 1708.06401v2*.

[Schoenberg et al., 2017] Schoenberg, F., Hoffmann, M., and Harrigan, R. (2017). A recursive point process model for infectious diseases. *arXiv 1703.08202v1*.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

[Yang et al., 2017] Yang, S. Y., Liu, A., Chen, J., and Hawkes, A. (2017). Applications of a multivariate hawkes process to joint modeling of sentiment and market return events. *Quantitative finance*, 18:295–310.

[Yu and Kak, 2012] Yu, S. and Kak, S. C. (2012). A survey of prediction using social media. *arXiv 1203.1647v1*, abs/1203.1647.