Inference for Hawkes process with Missing Data

Project by

Chandini Shetty

For the Partial Fulfillment
of the Requirements for the M.S. Degree
in Computer Engineering

University of California, Riverside
June 2017

Approved By:

Professor Christian R Shelton, Advisor
Department of Computer Science and Engineering
University of California,Riverside

Professor Nael Abu-Ghazaleh,
Department of Computer Science and Electrical Engineering
University of California, Riverside

Contents

List of Figures

Abstract

In this project, we look at probabilistic modeling of data with a Hawkes process, a stochastic process that belongs to the general class of multivariate Point Processes. Most often, real world data contains latent structure, which cannot be inferred directly and may instead require positing a latent variable model. Also sometimes the data is unobserved or unrecorded for certain interval of time and learning from such high dimensional data is a computational challenge. This work looks at fitting a Hawkes process and estimation of the parameters from data that includes such unobserved intervals by utilizing an Auxiliary Gibbs sampler and the Expectation-Maximization (EM) algorithm. We use the algorithm on a dataset from Chicago Police Department that contains homicide data for 30 years. We demonstrate the efficiency of the model in discovering clusters present in the data, by using latent variables.

Acknowledgement

I would like to express my deepest gratitude to my Research Advisor, Prof. Christian Shelton, for his unending patience and encouragement, that helped me complete this work. I got introduced to the field of machine learning and conducting academic research, through his coursework and project supervision. I am always grateful for everything that I have learned from him.

I would like to thank my Graduate Advisor, Prof. Nael Abu-Ghazaleh, for being part of my project presentation committee and in assisting me on all other aspects of my academic program, enabling my success. I would also like to thank all the members of R-LAIR for the weekly discussions and knowledge shared by them, in their respective areas of research.

Chapter 1

Introduction

Certain event data have stochastic excitation effects. The probability of their occurrence is increased by occurrence of previous events. Data that exhibits such properties include, for example, occurrences of gang-violence, earthquakes and social networks. An event at a certain point causes more events to be generated, which leads to a cascading effect. After an earthquake, the possibility of more earthquakes in nearby locations will increase significantly, due to the aftershocks triggering them. Similarly, crime data exhibits this type of contagious effects. A shooting between two rival gang factions will cause more shootings to follow because of retaliatory

effects. Regions that the gangs belong to also represent important aspects of these events.

Interactions between one event to another can exist if occurrence of one increases the probability of occurrence of the second. It is useful to discover interesting and hidden patterns that such event data may contain as it can help predict the occurrence of a earthquake or the possibility of breakout of gang violence in future.

An important consideration for such type of real-world events is that not all the information is recorded. There are events or random intervals for which information is unrecorded or unavailable. This could be systematic omission or missing-at-random. We demonstrate in this work that despite such missing data, we are able to recover meaningful structure from the data, by hypothesizing about latent structure.While the actual values of such hypothesized data may not be crucial, it could prove to be very crucial in developing a good model.

Event data, where we consider the times of the event occurrences as a set of events constituting a history or filtration on time, have been modeled with different stochastic processes. A Poisson process can model this with the assumption that events occur independent of each other. While, with a inhomogeneous Poisson process we assume the intensity at which the events occur is continuous and varying and events are still independent. These processes are inherently well suited to data that satisfies the independence assumption, i.e. they have been generated from an identical distribution, independently. To model the dependencies that exists in the times of the event data and their associated contextual information, more complex models have to be considered.

A Hawkes process, introduced by [**Hawkes**] , belongs to the general class of multivariate point process and has been widely studied for modeling data that exhibits excitatory effects. It has been extensively used to model financial data related to trade order and stock, with earthquake data to model the aftershock effects and so on. In this work we look at fitting a multivariate Hawkes process to gang-related homicide data obtained from Chicago Police department. We use the exponential and power kernels for the Hawkes Process in this work. We determine the parameters and the underlying latent structure by the use of latent variables and show that meaningful structure can be obtained from a simple Hawkes process with the use of auxiliary variables.

Chapter 2

Hawkes Process

2.1 Multivariate Hawkes Process

This section covers relevant background on the Hawkes process and its multivariate representation. The Hawkes process [**Hawkes**] also referred to as a self-exciting

point process belongs to the class of point processes that are continuous-time. It has been extensively used in data that has temporal dependencies between events, such as earthquakes after shock, in financial domains, predicting terrorist activities and crime. Most previous works in this area have considered univariate Hawkes process or assumed complete data. We specifically look at the multivariate Hawkes process along with a spatial component added, since this is more applicable to data that has possible interactions among the different labels or dimensions.

2.1.1 Hawkes Model representation

A multivariate Hawkes process [**Hawkes**] is a Hawkes process that consists of more than one component process. Throughout the course of this work, the constituent processes are referred to as labels. An event occurring in one label can cause events to occur in other labels and hence cause changes in the rate of events in those labels.

It is mathematically represented by its intensity function $\lambda(t)$ as

$$\lambda_l(t, h_t) = \mu_l + \sum_{i|t_i<t} \phi_{l_i,l}(t - t_i) \tag{2.1}$$

Here μ_l represents the base rate of events occurring in label l and $\phi(t)$ is the kernel function. $\phi_{l,l'}(t)$ represents the increase in the rate of label l' due to an event occurring in label l, t time units ago. The set of all events that occurred before time t, are denoted as $\mathcal{I}_t = \{i|t_i < t\}$ If we add a label $l = 0$, and set $\phi_{l,0} = 0$ and $\phi_{0,l} = \mu_l$, the expression for intensity function simplifies to contain the only the second term and base rate term is incorporated into it. This simplifies the notation and we now denote this set of event indices as $\mathcal{I}_t^0$, the set $\mathcal{I}_t$ plus the index of this special event.

2.1.2 Kernels for Hawkes process

The function $\phi_{l,l'}(t)$ is called the kernel function and models the temporal interdependencies between various event times that are observed as part of data to be studied. It can be factorized as product of a LxL matrix W, where L is the number of labels and $\phi(t)$, a base kernel function. $\phi_{l,l'}(t) = W_{l,l'}\phi(t)$. We have explored two possible definitions for the kernel functions in this work- the exponential kernel function $\phi(t) = e^{-\beta t}$ and a power kernel function $\phi(t) = (t+\gamma)^{-(1+\beta)}$. These kernels are parameterized by $\beta > 0$ and $\gamma > 0$.

An important requirement for the kernel function parameters for the Hawkes process is that the following condition should be satisfied, $\lambda \int_0^\infty \phi(t)dt < 1$, that is the spectral radius should be less than 1. [2] provide a mathematical proof for this conditions. Additionally they also explain stationarity requirements for the Hawkes process. Otherwise the process would not be considered stable and may grow infinitely in time i.e by generating a infinite number of events in a finite interval of time. This would invalidate the mathematical tractability of the Hawkes process and render it unsuitable for the data that is being considered for modeling here. Given this, the likelihood for the Hawkes process can be written as

$$p(x) = \exp(-\sum_i \Phi_{l_i,*}(T - t_i)) \prod_i \sum_{j \in \mathcal{I}_t^0} \phi_{l_j,l_i}(t_i - t_j) \tag{2.2}$$

Here $\Phi_{l_i,*}(T - t_i) = \sum_{l'} \Phi_{l_i,l'}(t)$ and $\Phi_{l_i,l'}(t) = \int_0^t \phi_{l_i,l'}(s)ds$.

With this likelihood formulation for multivariate Hawkes Process, the Maximum likelihood estimation(MLE) will result in determining the set of parameters that maximizes this quantity.

In addition it is important to note that exponential kernel for Hawkes process has Markovian property, as shown by [2]. However, the non-exponential kernel forms like power-kernel cannot be considered Markovian in nature, and the entire past history needs to be utilized for exact estimation.

Chapter 3

Latent Variable Models and Auxiliary Gibbs Sampling

This section provides relevant background on latent variable models and explains the auxiliary Gibbs sampling method for inference. It also describes the Expectation-maximization (EM) algorithm that is used to learn the model parameters.

3.1 Latent variable models

Latent variable models have been widely used in the analysis of document topics (LDA), speech recognition ,modeling social network and bio-informatics [5],[4]. Latent variable models try to recover latent structure from data that has large dimensions and can potentially discover hidden network or cluster structures. Such patterns in the data may not be easily discovered when the data is large or has high-dimensionality. By developing a latent variable model for such data we try to get a better understanding of the distribution of the observed events by positing the presence of extra latent variables. In certain types of data, it is necessary to make the assumption of missing data, for example in social network data. There could be missing edges,nodes or possibly both. Discovering good structure in the underlying network will involve making assumptions about this data.

Given some observed variable x, which is the evidence, we are interested in learning the probability distribution of x, $P(x)$. In a latent variable model, we augment x, with some unobserved variable z and instead try to reason about the joint distribution $P(x, z)$. Given this, it is possible to obtain $P(x)$ by marginalizing over the variables z , $P(x) = \sum_z P(x, z)$

For the Hawkes process, assuming the evidence/observation in our case event occurrence times are denoted as $\mathbf{x}$. The latent variable model will augment the evidence with unobserved events - denoted by variable z. This would allow us to better understand the joint distribution P(x,z). These latent variables in the model corresponds to adding new labels that can help explain the event times that were observed. Learning the model and its parameters will make inference possible, while allowing to predict missing event times or future event times.

3.2 Monte Carlo Methods

Markov chain Monte Carlo (MCMC) methods [13], such as the Metropolis-Hastings[14] algorithm and the Gibbs Sampling[9] are some of the commonly used and popular tools for the analysis of complex statistical models. In machine learning, MCMC methods are frequently applied to problems of integration and optimization, that

are in high-dimensions. In inference tasks, they have been used to estimate the posterior distributions for the unknown details of the statistical model. These could be the parameters of the model, latent variables, missing data etc. MCMC methods provide a strategy for generating samples from a distribution of interest using a Markov chain mechanism.

The algorithms generate samples, $x(i)$ that represent the state space of some target distribution X whose stationary distribution is $p(x)$ by building a Markov chain. For the purpose of the multivariate Hawkes model, we use a variation of the MH-Gibbs sampling, which is known as the auxiliary Gibbs Sampler, which is example of Reversible jump MCMC (RJMCMC). RJMCMC methods are more frequently used to solve the problem of model selection. Several works have demonstrated the use of RJMCMC methods for problems such as components in mixture models [17], neuron estimation in neural networks [11],[15], [1] where the sampling helps to chose from a family of models

For the purposes of the problem that is modeled here we use an adaptation of a particular implementation of the RJMCMC sampler - the Auxiliary Gibbs Sampler [16] to perform inference. This sampler has been used in earlier works for inferencing task in PCIM models [10] for inferencing.

Specifically, two sets of auxiliary variables are used with our sampler. The first set represents the parent-child relation or the branching structure of the multivariate Hawkes process. If the parent events are denoted by $a = a_1, a_2...a_n$, where a_i is the parent event of event with index i. Then $p(x, a)$ would be the joint distribution and whose marginal, p(x) is the prior distribution of the events in the set x. The joint distribution $p(x, a)$ is now given by.

$$p(x,a) = \prod_i \phi_{l_{a_i},l_i}(t_i - t_{a_i}) \exp(-\bar{\phi}_{l_i,*}(T - t_i)) \tag{3.1}$$

The second set of auxiliary variables are extra events introduced to the observed event sequence that is available. Such events are referred to as virtual events, since they are not real and instead represent possible new events in certain unobserved intervals. This idea has been used by [16] for PCIM inferencing, which is also non-Markovian. like the multivariate Hawkes process. Virtual events are the possibilities for children events for the real events and are generated from a Poisson process with rate $\kappa * \phi_{l_i,l}(t - t_i)$. Using this as the rate, the sampler generates many possible children events for the real events, including the root event. Let $\tilde{x}$ denote the set of virtual event, $\tilde{x} = \{(\tilde{t}_1, \tilde{l}_1), (\tilde{t}_2, \tilde{l}_2)....(\tilde{t}_n, \tilde{l}_n)\}$

The augmented joint distribution is denoted as $p(x, a)$, where a denotes the auxiliary variables. the sampler has three possible transition moves over this distribution. 1) Re-sampling the virtual children (2) transforming a virtual event to real (or vice-versa) (3) altering the parent-child relationships. Using the auxiliary variables and the moves described, the new joint distribution will be over $p(x, a, \tilde{x}, \tilde{a}$ and has the same marginal $p(x)$ as the original process.

3.3 EM Algorithm

The Expectation maximization (EM) [8] algorithm is a popular algorithm for parameter estimation with maximum likelihood , in latent variable model. For the Hawkes process, it is particularly useful in determing the latent structure present amongst the different labels and can lead to discovering clustering structure amongst labels. We use this technique to learn the parameters of the model which consists of two alternating steps: the E-step and the M-step. In particular, we use the Monte Carlo EM (MCEM) algorithm that employs a MCMC sampler [7] for expectation estimates.

E-Step: In the E-step or the Expectation step, we generate samples from the posterior distribution over latent variables, given the evidence events. A sample of event sequences or trajectories are generated with potential parent-child $(t_{a_i} \rightarrow t_i)$ relations. The Hawkes process to begin with is initialized with random values for the parameters μ_l, β, γ, W. We set these values appropriately by drawing from random uniform/real valued distributions. Given the parameters of the Hawkes process and our evidence data, the sampler is initialized and we run it for some number of iterations to obtain samples. This constitutes the E-Step.

M-step: Given the probabilistic samples obtained in E-Step, from the posterior distribution we are interested in, the M-step involves computing the maximum likelihood given the sampled data. This results in calculating the new parameters for the data by maximizing the sum of the log-likelihood, with addition of a L1 regularization for the samples. This gives new values $\mu_l', \beta', \gamma', W'$. The Hawkes process is now initialized with the new parameters and we repeat the E-step.

The EM algorithm is repeated for some desired number of iterations, until we see convergence in the parameters. As the computation progress, the number of possible parent-child relations generated reduces and there are fewer such moves detected. This results in the parameters stabilizing around certain ranges of values.

Maximizing the log-likelihood, for this representation, is not trivial and generic

numerical optimizations techniques may not directly apply. In the next section, we show the method of computing these using closed form analytical expression and simple optimization for calculating the kernel parameters is described.

Chapter 4

Parameteric Estimation with EM

The previous chapter described the EM algorithm mechanism for learning the parameters. This chapter will go over the optimization procedure used in estimating the necessary parameters, as described in the M-step

4.1 Parameter Estimation

The joint distribution over the evidence x and auxiliary variable a is given by

$$p(x,a) = \prod_i \phi_{l_{a_i},l_i}(t_i - t_{a_i}) \exp(-\Phi_{l_i,*}(T - t_i)) \tag{4.1}$$

The log-likelihood is given by

$$\mathcal{L}(x) = -\sum_i \Phi_{l_i,*}(T - t_i) + \sum_i \log(\sum_{j|t_j<t_i} \phi_{l_j,l_i}(t_i - t_j) + \mu_{l_i}) \tag{4.2}$$

But, if we consider the auxiliary variable based formulation of equation 5.1 the log-likelihood is given by

$$\mathcal{L}(x,a) = -\sum_i \Phi_{l_i,*}(T - t_i) + \sum_i \log(\sum_{i|l_{a_i}\neq -1} \phi_{l_{a_i},l_i}(t_i - t_{a_i})) \tag{4.3}$$

$$+ \sum_{i|a_i=-1} \log(\mu_{l_i}) - \sum_l \mu_l * T + \lambda||W||_1$$

This gives the likelihood estimation over the auxiliary variables and is computationally easier, since the number of events that would be considered in estimating this is a smaller subset than the originally considered likelihood. We obtain this Equation (4.3) by separating out events whose parent is the root event, $l_{a_i} = -1$

and events that have non-root parents, i.e $l_{a_i} \neq -1$, as is the case with the second term. The term with $\log(\mu_{l_i})$ accounts for the contribution by events whose parent label is -1. We also add an L1-regularization term to achieve sparsity in the weights, because each label interacts with only few other labels.

Further, this formulation gives a closed form for estimating the W using analytical methods instead of numerical optimization, while the base rate calculation can be done by averaging the number of background events in each label over the total time. This reduces the scope of optimization to just the kernel parameters, β for the exponential kernel an β and γ for the power kernel.

4.1.1 Estimating the W parameters

The matrix W is an LxL matrix where L is the number of labels. Each element w_{ij} represents the influence of the i'th label on the j'th label. If we fix the kernel parameters and base rates, and differentiate Equation (4.3) with respect to w_{ij}, and set it to zero, we can solve for w_{ij} analytically and we no longer need to optimize this using numerical methods. The derivation for this is presented below. Rewriting the terms of Equation 5.3 to contain the $w_{i,j}$ terms:

$$\Phi_{l_i,*}(T - t_i) = \sum_l \Phi_{l_i,l_j}(T - t_i) = \sum_{l_i} w_{l_i,l_j}\phi(T - t_i)$$

$$\phi_{l_{a_i},l_i} = w_{l_{a_i},l_i}\phi(t_i - t_{a_i})$$

So Equation 4.3 is now rewritten as

$$\mathcal{L}(x,a) = -\sum_i \sum_l w_{l_i,l_j}\phi(T - t_i) + \sum_i \log(\sum_{i|l_{a_i}\neq -1} w_{l_{a_i},l_i}\phi(t_i - t_{a_i}) + \sum_{i|a_i=-1} \log(\mu_{l_i})$$

$$-\sum_{l}\mu_l * T + \lambda * \sum_{l_i,l_j} w_{l_i,l_j}$$

Taking the partial derivative with respect to each w_{l_i,l_j} gives

$$\frac{\partial \mathcal{L}}{\partial w_{l_i,l_j}} = 0$$

$$w_{l_i,l_j} = \frac{n_{i,j}}{\sum_{t_k|l_k=l_i} \phi(T-t_k) + \lambda} \tag{4.4}$$

here $n_{i,j}$ is the number of events with label l_i as parent and label l_j as the children events label. The denominator term is sum over all of the events with time t_k that has the same label as t_i

Once the W matrix is calculated with the above set of Equations, we next estimate the the background rates, by obtaining the derivative of $\mathcal{L}$ with respect to μ_l, w as

$$\frac{\partial \mathcal{L}}{\partial \mu_l} = 0$$

$$\mu_l = \frac{n_l}{T}$$

where n_l is the number of background events of label l.

Next, we can optimize for the kernel parameters which will be a one variable optimization (β) for exponential kernel and two-variable (β and γ) for the power kernel. We use standard library optimization algorithms for this low dimensional optimization problem.

Chapter 5

Fitting the Model to Real Data

In this section, we look at applying the algorithm and methods developed in previous chapters, to a real world dataset. We experiment with both a exponential and power kernel and summarize the results here with visualizations.

5.1 Dataset Description

We use a real dataset obtained from the Chicago Police Department. The Chicago Homicide Data(CHD), owned and distributed by ICPSR, contains information on every homicide that occurred between 1965 – 1995 in the city of Chicago. It contains both victim-level and offender-level data files with about 23,817 records for victims

and 26,030 records for offenders. The numbers differ since there can be more than one victim per offender and vice versa. In addition, the offender level data contains only records where at least one offender was known to the police. Any unsolved homicides will appear only in victim-level files and omitted from offender files. The CHD data is completely geo-coded, but the exact location of the incident has been anonymized for confidentiality purposes. However each address has been instead mapped to one of the 77 community areas of the city of Chicago.

The data set has multiple predictor variable for each homicide such as the racial/ethnic group of all victims and offenders, offender-victims relation, type of weapon used, injury data. For the purposes of our experiment, we use crimes coded as homicides related to gang members- that is either the victim or the offender was affiliated with Chicago's street gangs. The causal factor code $= 140$ indicates this. The time of the incident is converted as a fraction of the days.

After this processing, we have 2196 events related to gang-related homicide, with each event assigned to a corresponding label 1–77, which is one of the 77 community areas of the city of Chicago. That start time is 0 and end time corresponds to 11322 in days. This data is then parsed by the code to build a trajectory object for the multi-variate Hawkes process. Each dimension of the Hawkes process represents a label that has value in the range 1–77 of the community areas. For every label/region, the corresponding event times of the homicides form the set of times associated with that label.

Using these semantics for the trajectory information, we use the sampling algorithm to generate the trajectories for intervals that are not observed and generate possible parent child relation between the events as described in the section on Auxiliary Gibbs Sampler.

We look at running the algorithm for two variations of the Hawkes kernel—the exponential kernel and power kernel and consider two possible cases within each.

5.2 Results for Exponential Kernel

5.2.1 Without latent variables

In this experiment, we fit the model for the raw data with an exponential kernel with no latent variables considered. Specifically we look at this as a 77-spatial labels Hawkes process, with the data set as the event times for every homicide that occurs in a particular label or community area. Given this formulation, we can do a maximum likelihood estimation for the obtaining the parameters of the models- we are interested in learning the background rate of crime in each of the labels, μ_l, the kernel parameter β, and the matrix W which determines the network structure and excitations between the labels of the Hawkes process.

We run this experiment with a initial burn-in of 50000 samples and then continue for 150-200 additional iterations. In each iteration, we initialize the auxiliary Gibbs Sampler with trajectory data, and obtain samples of sizes 1000. We choose 1000 after experimenting with different values, since this gives good convergence progress after each iteration without taking too long to compute. Next we estimate the parameters by maximizing the total log-likelihood of the samples. This is same as the general MLE estimation procedure for determining parameters of the model given data and the sampling tries to infer the latent relation amongst the different labels. Once the parameters are obtained, the kernel is initialized with the new values of μ_l, β and W and we repeat the iteration. This is done several times, with a range of value for lambda- the regularization parameter in the range 1e2 to 1e7 and until the parameter values converge.

Figure 5.1 shows the results for the exponential kernel, with values obtained the kernel parameter as $\beta = 0.0092$ and lambda value of 1e5. The background rates are plotted corresponding to the different communities as a choropleth graph on a map of city pf Chicago. To visualize the matrix, we plot it as graph using Gephi[3] and use the Modularity Class algorithm [6] for community detection. Studying the graph reveals a clustering structure with several communities overall and no clear clusters. The modularity class returns about 15 communities, with some having 2 nodes per cluster.

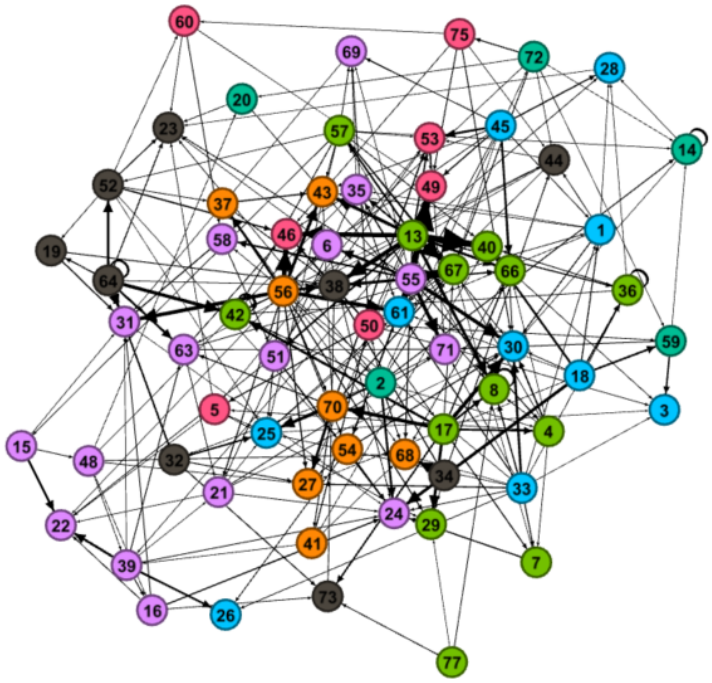


Figure 5.1: Network structure for the exponential kernel Hawkes process with no latent variables

5.2.2 Exponential Kernel With Latent Variables

For this experiment, we add three latent variables to the model. These are denoted as Node 78-LV1,Node 79-LV2 and Node 80 LV3 in the data and figures. We repeat the experiment as above and estimate the kernel parameter as $\beta = 0.0137$ with lambda value of 1e5. The corresponding network graph structure is shown in Figure 5.2 and associated regions are shown in Figure 5.3. The estimated background rates are plotted in Figure 5.4 and 5.5. In addition we are interested in knowing the influence of the latent variable and the mutual interactions in between them and other regions. This is plotted in All of the W parameter structures are visualized using a network graph visualization tool. The noes in the graph are sized according to their degree- higher the degree, larger the node size. Graph reveals a clustering structure with 5 communities overall and the member nodes/label are of the same color for a particular cluster. The three latent variables have significant number of interactions and trigger possible events in these regions. The thickness of the edge and direction of the edge depict the cause and effect between the latent variable nodes and other nodes.

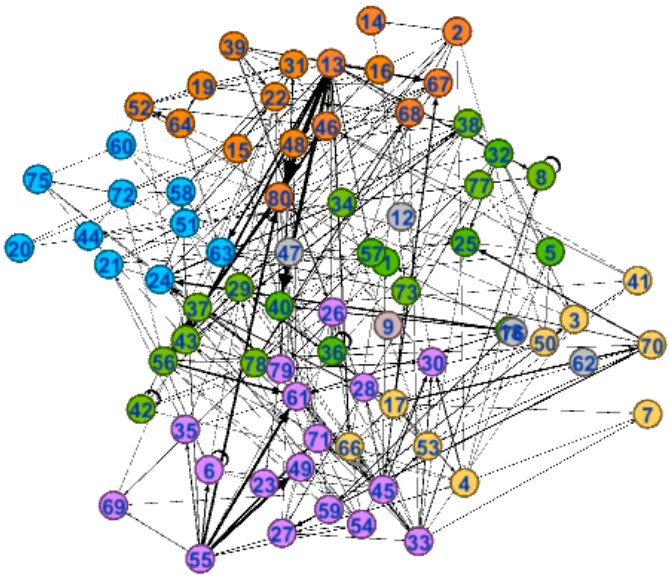


Figure 5.2: Network structure for the Exponential kernel Hawkes process with 3 latent variables

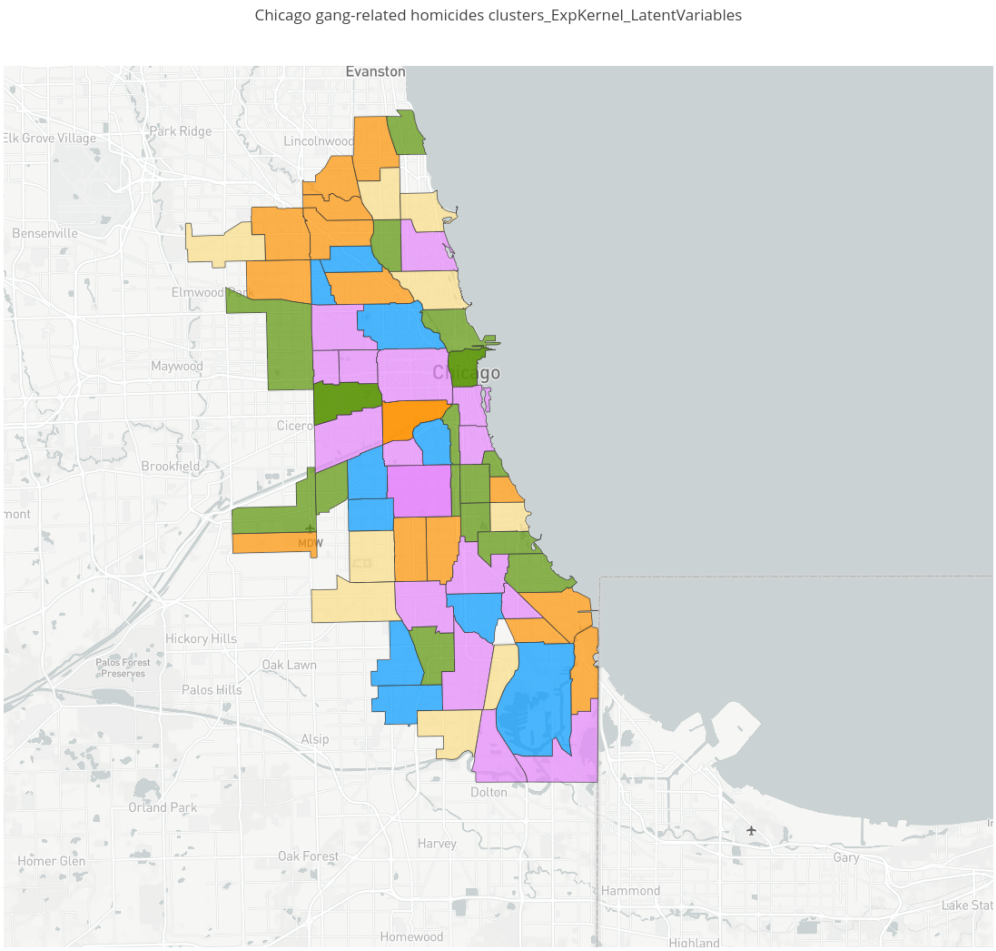
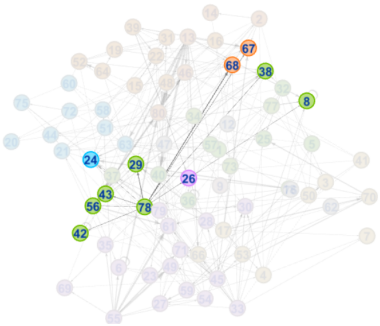
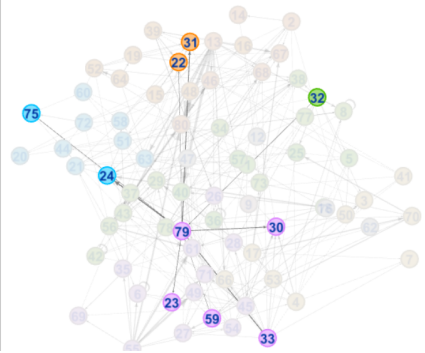


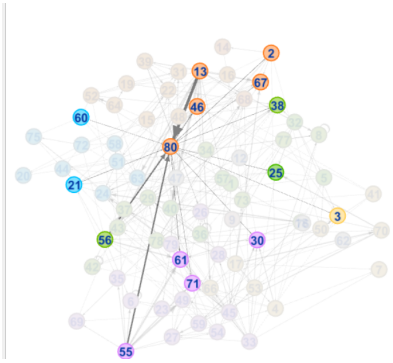
Figure 5.3: Geo visualization of the regions corresponding the clusters of Fig:5.2



(a) Interactions for latent variable 1- Node 78



(b) Interactions for latent variable 2- Node 79



(c) Interactions for latent variable 3- Node 80

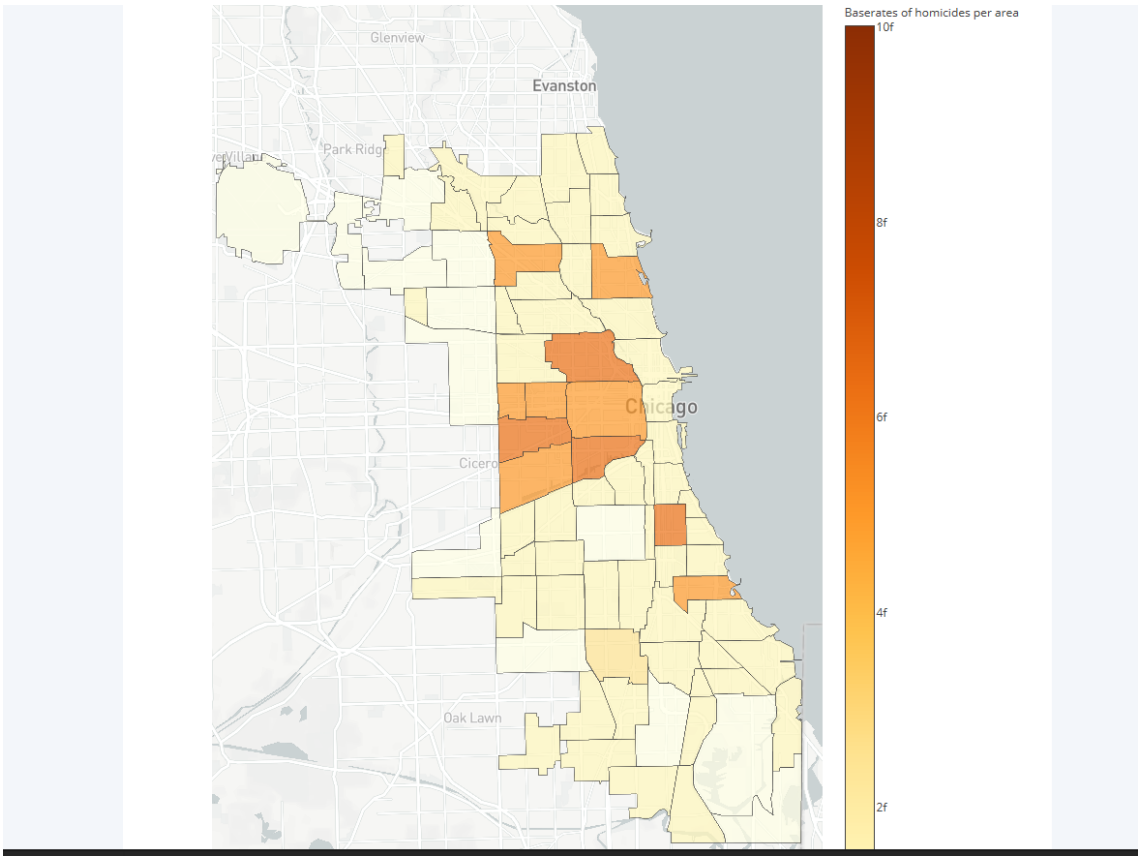


Figure 5.4: Baserates(μ_l) estimated for the Exponential kernel Hawkes process with no latent variables

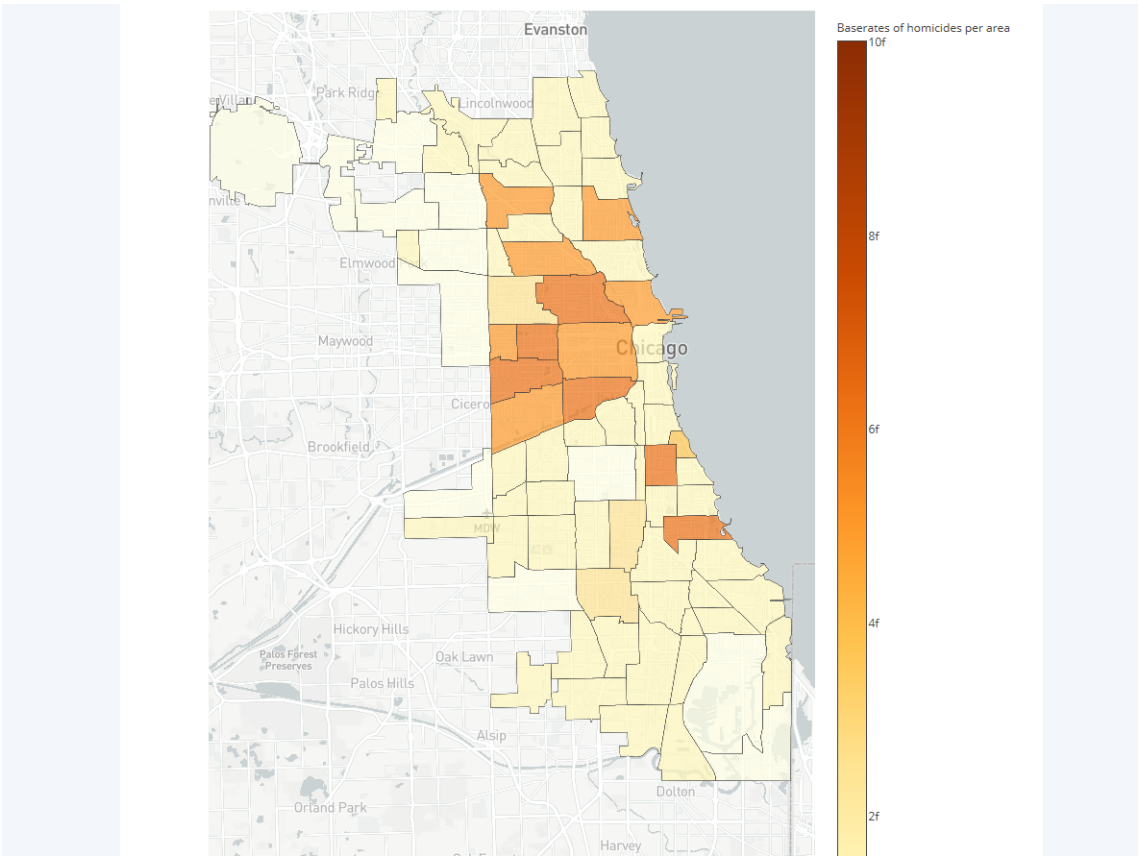


Figure 5.5: Baserates(μ_l) estimated for the Exponential kernel Hawkes process with 3 latent variables

5.3 Results for Power Kernel

In this section we repeat the experiments done in previous section, with a power kernel and summarize the results.

5.3.1 Without Latent variables

As is the case with the exponential kernel, we repeat the case with no latent variables added. The kernel parameters stabilizes to around $\beta = 1.135$ and $\gamma = 57.4062$. We run burn-in for 50000 iterations and use sample size 1000 for the MLE. The network structure is shown in Figure 5.6 with regularization set to 1e4. The modularity class algorithm run through Gephi returns about 10 communities.

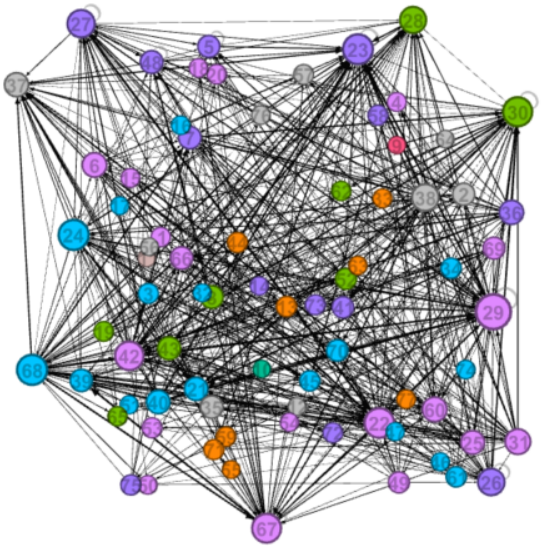


Figure 5.6: Network structure for the Power kernel Hawkes process with no latent variables added

5.3.2 Power kernel with Latent variables

Here we add 3 latent variables to the Hawkes process model just as in the case of the exponential kernel. These are denoted as Node 78-LV1,Node 79-LV2 and Node 80-LV3 in the data and figures. We estimate the kernel parameter as approximately $beta = 1.34937$ and $gamma = 60.2072$. Also, the networks graphs with interactions of the latent variable nodes with other nodes is shown in Figure 5.7- 5.8. After the running the modularity class algorithms, we detect 4 community cluster depicted in magenta, green, blue and black. Also shown in the 3 subplots are the interactions for the three latent variables with other nodes. Figure 5.9 shows the background rate with the latent variables added for the power kernel.

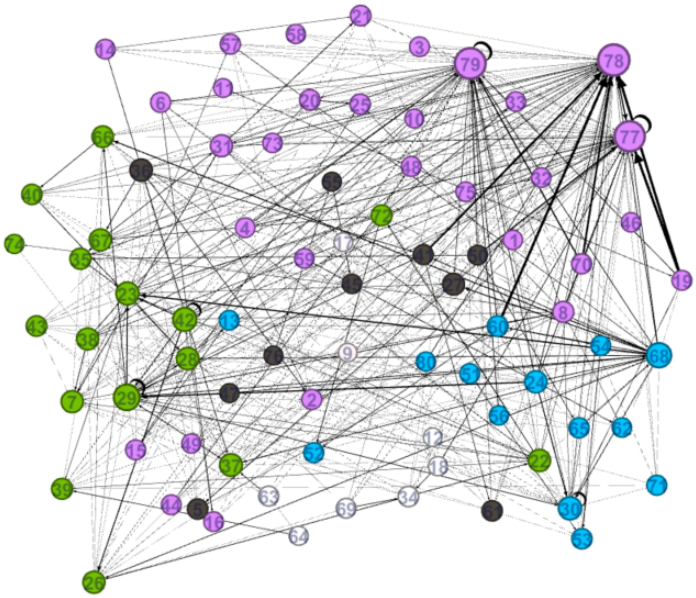


Figure 5.7: Network structure for the Power kernel Hawkes process with 3 latent variables

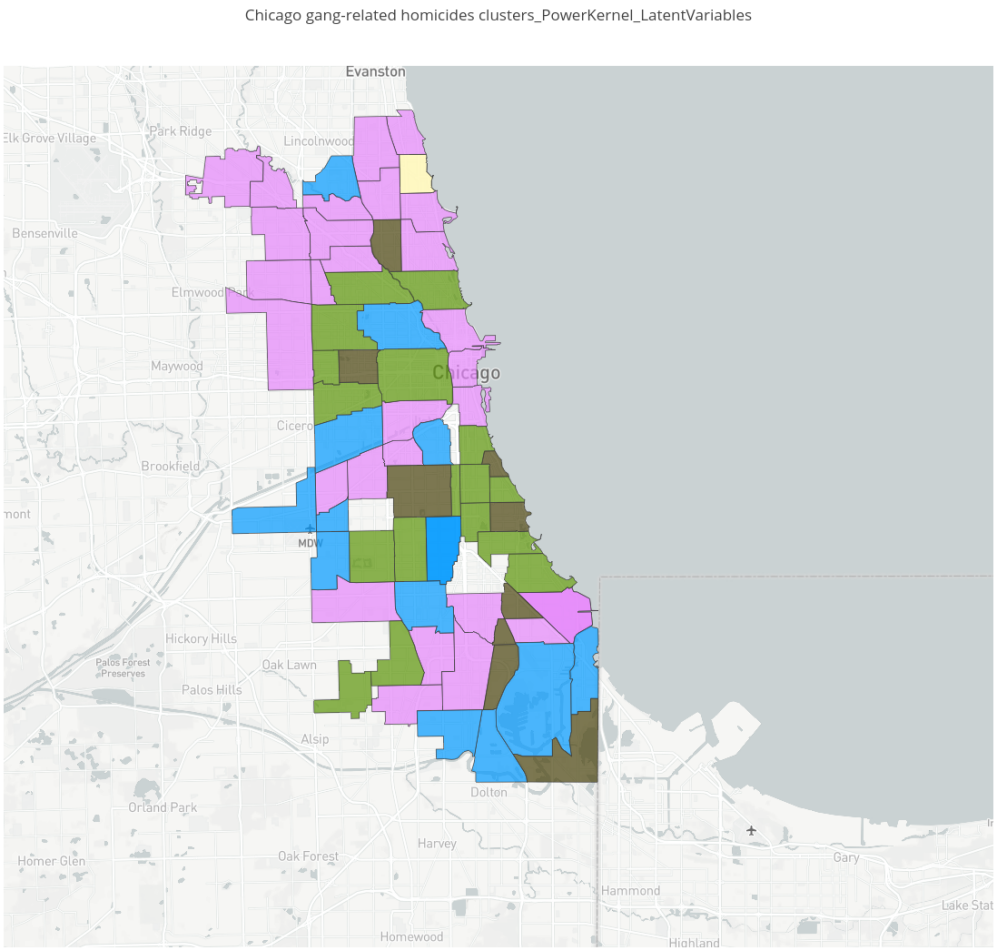
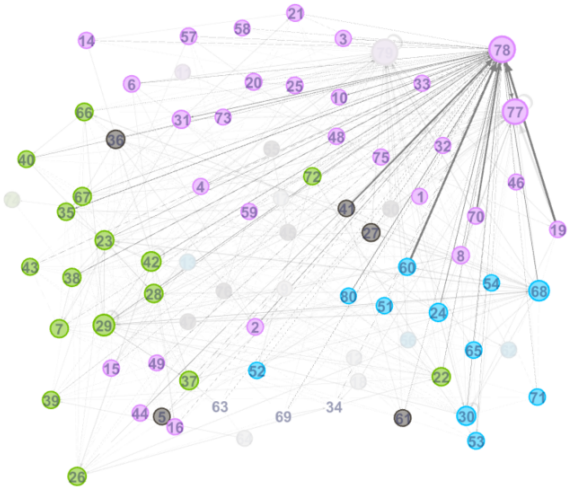
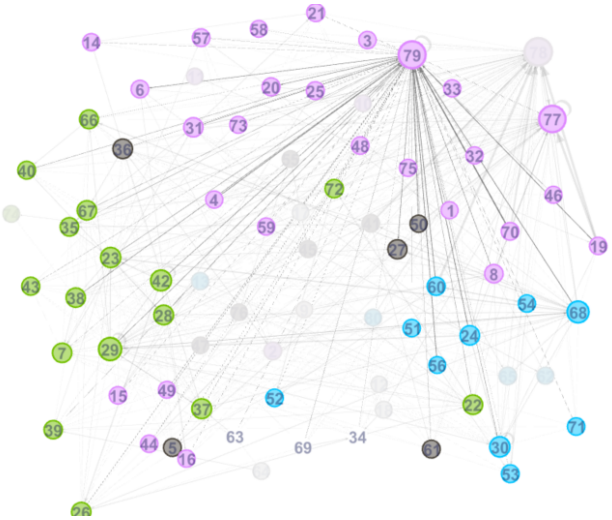


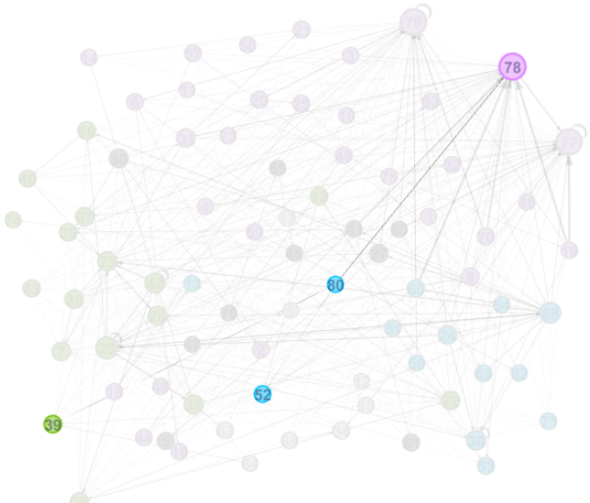
Figure 5.8: Geo visualization of the regions corresponding the clusters of Fig:5.7



(a) Interactions for latent variable 1- Node 78



(b) Interactions for latent variable 2- Node 79



(c) Interactions for latent variable 3- Node 80

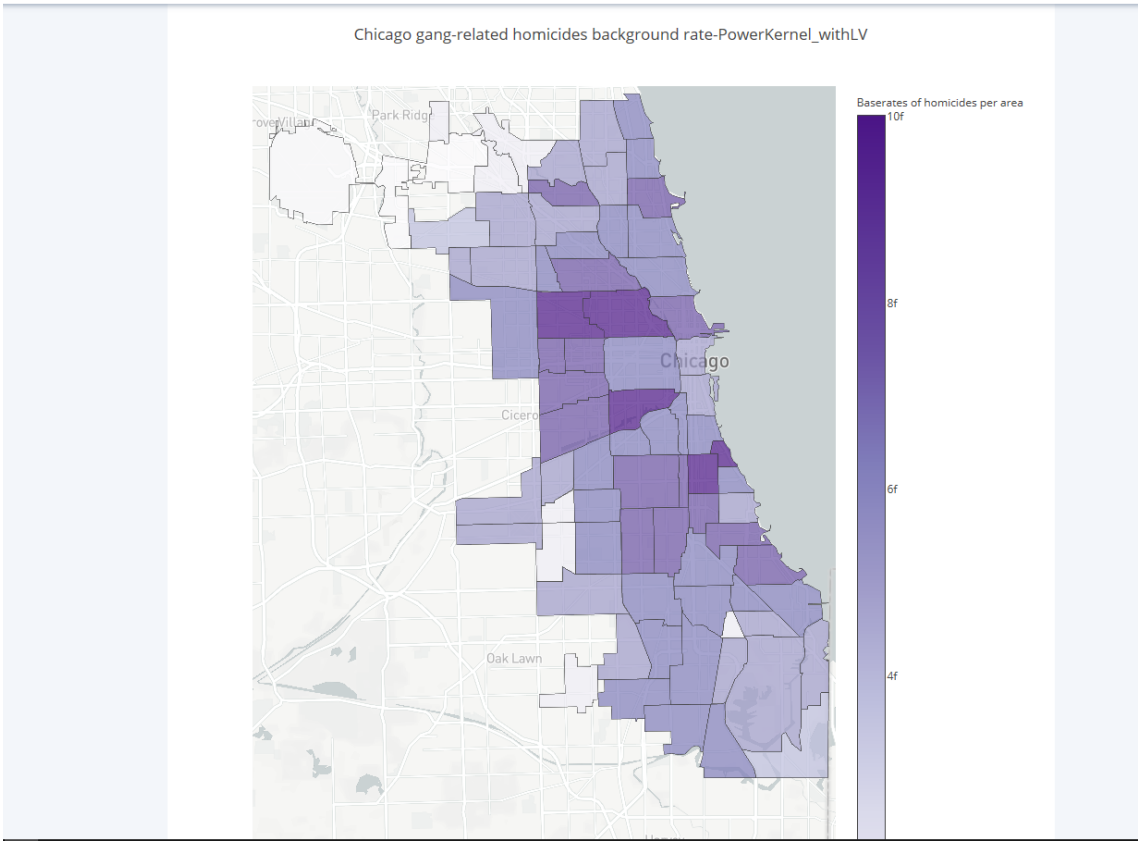


Figure 5.9: Baserates(μ_l) estimated for the Power kernel Hawkes process with 3 latent variables

Chapter 6

Conclusion

From the results of chapter 5, we can draw few interesting conclusions. With the addition of latent variables that model the unobserved events we obtain a better cluster structure amongst the labels. In contrast, the network structure which is obtained with the plain Hawkes process does not produce any clear clustering of the communities.

The power kernel appears to model the data much better than the exponential kernel, in terms of giving fewer clusters. Similar results have been also obtained by others who have modeled the same data with more complex Hawkes process [12]. The results showing the background rates μ_l, for the exponential and power kernel, shows different levels of crime rate occurring in the regions. The power kernel results seems to indicate a higher level of activity in South and South-west regions of the city of Chicago.

We achieve good results for the model and with a simpler formulation of the Hawkes process and capture the network community structure with the use of latent variables.

Bibliography

[1] Christophe Andrieu, Nando De Freitas, and Arnaud Doucet. "Robust full Bayesian learning for radial basis networks". In: *Neural Computation* 13.10 (2001), pp. 2359–2407.

[2] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. "Hawkes processes in finance". In: *Market Microstructure and Liquidity* 1.01 (2015), p. 1550005.

[3] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: (2009). URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[4] David M Blei. "Build, compute, critique, repeat: Data analysis with latent variable models". In: *Annual Review of Statistics and Its Application* 1 (2014), pp. 203–232.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[6] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[7] James G Booth and James P Hobert. "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.1 (1999), pp. 265–285.

[8] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)* (1977), pp. 1–38.

[9] Stuart Geman and Donald Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), pp. 721–741.

[10] Asela Gunawardana, Christopher Meek, and Puyang Xu. "A model for temporal dependencies in event streams". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1962–1970.

[11] CC Holmes and BK Mallick. “Bayesian radial basis functions of variable dimension”. In: *Neural computation* 10.5 (1998), pp. 1217–1233.

[12] Scott Linderman and Ryan Adams. “Discovering Latent Network Structure in Point Process Data”. In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, 2014, pp. 1413–1421. URL: http://proceedings.mlr.press/v32/linderman14.html.

[13] Nicholas Metropolis and Stanislaw Ulam. “The monte carlo method”. In: *Journal of the American statistical association* 44.247 (1949), pp. 335–341.

[14] Nicholas Metropolis et al. “Equation of state calculations by fast computing machines”. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[15] Peter Müller and David Rios Insua. “Issues in Bayesian analysis of neural network models”. In: *Neural Computation* 10.3 (1998), pp. 749–770.

[16] Zhen Qin and Christian R Shelton. “Auxiliary Gibbs Sampling for Inference in Piecewise-Constant Conditional Intensity Models.” In:

[17] Sylvia Richardson and Peter J Green. “On Bayesian analysis of mixtures with an unknown number of components (with discussion)”. In: *Journal of the*

Royal Statistical Society: series B (statistical methodology) 59.4 (1997), pp. 731–792.