Compression of spatio-temporal networks via point-to-point process models

Xiaoyue Li
Statistics Department
University of California, Davis
Davis, California 95616
xyrli@ucdavis.edu

James Sharpnack
Statistics Department
University of California, Davis
Davis, California 95616
jsharpna@ucdavis.edu

ABSTRACT

A point-to-point process describes a dynamic network where a set of edge events are observed, each of which is associated with a time of occurrence and two vertices lying in their state spaces. This study intends to investigate one application of such processes, using NYC Taxi and Limousine Commission dataset that reports taxi trips between two locations at a certain time. Here a point-to-point process is formed with edge events being taxi trips and the vertices adjacent to the edge events are pick-up and drop-off locations, described by latitude and longitude pairs. The intensity of an edge event can have a temporal dependence in addition to being dependent on a latent, spatially-coherent community structure for the vertices. To this end, we have developed a methodology that estimates a spatially smoothed community structure and localizes temporal changepoints for point-to-point processes. By applying this to our dataset, we can explore the spatio-temporal dynamics of the demand of taxi trips. More specifically, with reasonable assumptions, the latent community structure is estimated by spectral partitioning based on a low-rank reconstruction of aggregated taxi-trip network; and the temporal changepoint localization can be carried out by solving a matrix group fused LASSO program.

KEYWORDS

Point-to-point process, spatio-temporal networks, dynamic graph models

ACM Reference format:
Xiaoyue Li and James Sharpnack. 2017. Compression of spatio-temporal networks via point-to-point process models. In *Proceedings of International Workshop on Mining and Learning with Graphs, Halifax, Nova Scotia, Canada, August 2017 (WMLG),* 7 pages.
DOI:

1 INTRODUCTION

1.1 Taxi trip demand estimation

The increasing prevalence of ride-sharing and car-sharing services, and the advent of self-driving cars, necessitates the development of statistical tools for trip demand estimation. We will analyze the NYC Taxi and Limousine Commission dataset (http://nyc.gov) which reports taxi trips between two locations at a certain time (see Figure 1 for a depiction of the dataset). We would like to simultaneously explore the spatial and temporal effects on trip frequencies. A trip consists of the pick-up and drop-off location for the ride and the departure time, and we will assume that after conditioning on these effects the trip duration and other factors do not have a significant influence on the demand. Because we can think of a trip as forming a directed edge between the two locations, the trip dataset can be thought of as a dynamic network, because edges occur at random times. Because these edges are counting processes of trips from one point to another point, we will call this a *point-to-point process*. We will focus on the demand estimation procedures that provide compressed versions of the network, which is an important feature for mobile applications.

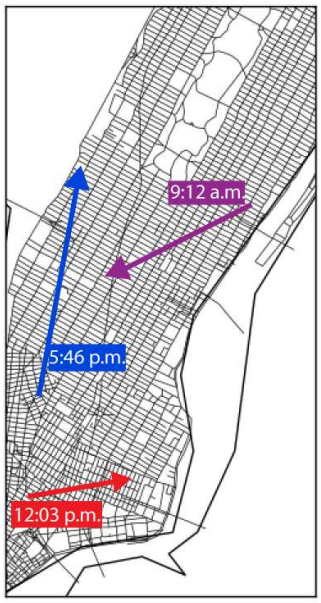

Figure 1: Three records in the NYC Taxicab Dataset, each consists of a timestamp and two locations.

As a preprocessing step we will aggregate the trips spatially and temporally, using a fine mesh, for example 1km x 1km regions spatially and 1 hour temporally. So, we store the number of trips between any two 1km^2 grid squares within any given hour as a tensor in memory. Consider the heatmap of trip counts leaving

WMLG, Halifax, Nova Scotia, Canada

DOI:

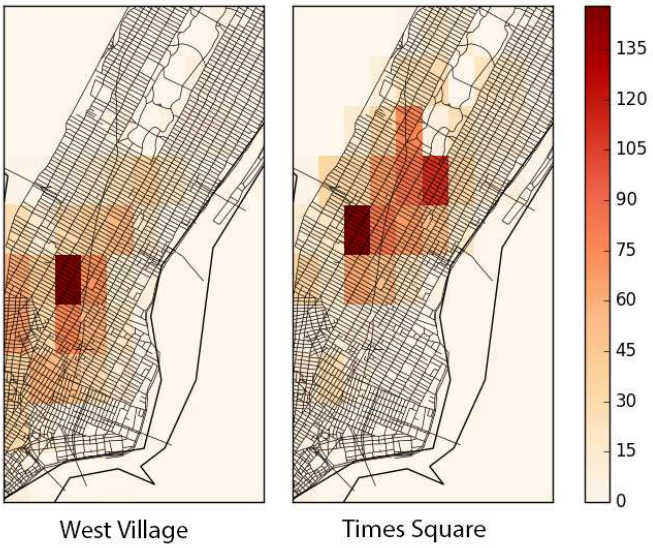

Figure 2: Heatmap of drop-off locations for taxi trips originating in the Times Square (right figure) and West Greenwich Village (left figure) from 12pm-4pm on 15 Jan. 2015.

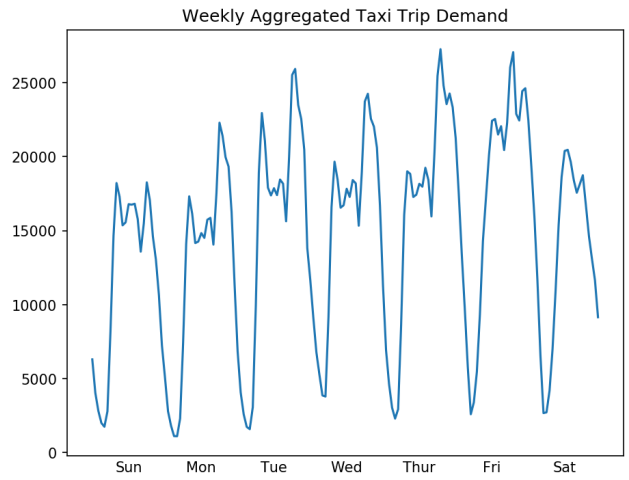

Figure 3: Trip counts within each hour throughout the week.

two different locations in Figure 2 aggregated over a 4 hour time period on 15 Jan. 2016 (a Thursday). Trips leaving Times Square have a significantly different distribution than those leaving West Greenwich Village. We expect that the nature of those locations (tourist destination, place of work, residence, etc.) and the specific time will interact to determine the trip demand. Thus, the locations are assumed to be members of communities of locations, i.e. the residential blocks, tourist areas, nightlife destinations, etc. This latent membership model persists across time, since a location does not stop being residential, but the demands between the communities will certainly change.

In order to account for weekly periodicity in the dataset, we further aggregate the trips by time from the beginning of the week. There are obvious changes in demand over time. If we just look at the overall number of trips over time of the week, we can see that the trip counts peak during the middle of the day, and then die down at night (see Figure 3). This constitutes a base rate, which we would like to account for separately in a semi-parametric fashion.

After the base rate has been removed, we would like to estimate time points across which we see large changes in relative demand. Because we are assuming that the intensities respect the community structure, we would like to estimate separately the demand between different communities change over time. To this end, we want to discover temporal changepoints, at which the relative demands between the communities change appreciably.

A changepoint model has a few advantages. It allows for interpretable results, in that the predicted demand between the West Village residential community and the Central Park museum community depends only on whether you are during early morning rush hour, afternoon on a weekend, etc. It is locally adaptive, in that if the demand changes significantly during a time interval, then there will be many changepoints, otherwise few changepoints will be detected. It also allows compress-ability of the output, so that a server might estimate the demand off-line and then the results will be sent to mobile devices by just relaying the temporal changepoints and the demands between the communities.

To summarize we want to model the trip demand by incorporating the following aspects described above:

(1) a location-based latent membership model dictates how the location affects the demand,
(2) a base rate is accounted for either as an offset or as a preprocessing step,
(3) temporal changepoints determine the time dependence of the trip demands.

For mobile applications, estimating the demand in this way has the advantage of compressing the estimated demand. Both modelling assumptions above lead to compressed demand functions. A spatial membership model means that only the location membership and the between membership demands need to be stored and transmitted. The temporal changepoint model means that the changepoint times and the change in demands across these times is maintained. The base rate is a scalar function, so it is assumed to require minimal memory.

1.2 Point-to-point processes

A point-to-point process consists of triples $(t, s_0, s_1) \in [0, \infty) \times \mathcal{X} \times \mathcal{X}$, which indicates the two locations in the state space $\mathcal{X}$ between which an interaction occurs and the time, t, at which this occurs. For the trip demand estimation, the state space $\mathcal{X} \subseteq \mathbb{R}^2$ which corresponds to the latitude and longitude of the locations. The basic operation that defines our stochastic process is counting. Given an interval of time $(t_0, t_1]$ and measurable sets $S_0, S_1 \subseteq \mathcal{X}$, we can count the number of trips from any point within S_0 to any point within S_1 between times t_0 and t_1. We will assume that this is a Poisson process in that the number of such trips is a Poisson random variable.

Let $N_t(S_0, S_1)$ denote the number of trips from a location within the set S_0 to the set S_1 before and including time t. A point-to-point process has *independent spatial increments* if for any measurable set $S' \subseteq \mathcal{X}$ and disjoint sets, $S_0, S_1, \ldots, S_k \subset \mathcal{X}$, then $\{N_t(S_i, S')\}_{i=1}^k$ are independent stochastic processes. It further has *independent temporal increments* if for disjoint time intervals, $[t_{i,0}, t_{i,1}]$ then $\{N_{t_{i,1}} - N_{t_{i,0}}\}_i$ are independent random measures. For taxi trips, we will assume that N_t is spatially and temporally inhomogeneous, in that if we take a measurable set, $S \subset \mathcal{X}$, and translate S (assuming there is such a group action on $\mathcal{X}$), then the distribution of the process, $N_t(S, \cdot)$, can change.

Let us discretize the continuous point-to-point process with a fine mesh to aggregate the taxi trips spatially and temporally. Suppose the mesh results in a grid of M_S rectangular sub-areas, $S_1, S_2, ..., S_{M_S}$, partitioning a rectangular region $\mathcal{X}$, and M_T time intervals $(t_q, t_q + \Delta t]_{q=1}^{M_T}$ of equal-length, $\Delta t = \frac{T}{M_T}$, partitioning the time of interest, $(0, T]$. Denote the observed counts within mesh elements as $A_{ij}^{(q)} = N_{t_q+\Delta t}(S_i, S_j) - N_{t_q}(S_i, S_j)$, which are assumed to be independent Poisson random variables. Then denote the trip intensity to be $\Lambda_{ij}^{(q)} = \mathbb{E}A_{ij}^{(q)}$.

We can extend the stochastic block model [24] to this setting by considering spatial block structure. We assume that there exists two different coarse partitions of the spatial mesh, $\{B_k^-\}_{k=1}^K$, $\{B_k^+\}_{k=1}^K$, which dictates the intensity of the point-to-point process. Specifically, the point-to-point process is *block structured* if the intensity function is only spatially dependent on which block the mesh element is a member of, i.e.

$$\Lambda_{ij}^{(q)} = \Lambda_{i'j'}^{(q)}, \quad \text{if } i, i' \in B_{k_0}^-, j, j' \in B_{k_1}^+$$

for some choice of departure block $B_{k_0}^-$ and arrival block $B_{k_1}^+$. The temporal dependence will be modelled through a base rate component that impacts all block pairs the same and a piecewise constant component. The base rate is a time-dependent scalar function, $\lambda_0^{(q)}$, that accounts for the cumulative temporal trend. After dividing by the base rate, we will assume that the intensity is piecewise constant, in that there are changepoints, $\{\tau_c\}_{c=1}^C$ such that

$$\frac{\Lambda_{ij}^{(q)}}{\lambda_0^{(q)}} = \frac{\Lambda_{ij}^{(q')}}{\lambda_0^{(q')}}, \quad \text{if } \tau_{c-1} \leq q \leq q' < \tau_c.$$

While these assumptions may or may not be an accurate modelling assumption, they will result in methodology that is produces compressed and interpret-able demand functions.

1.3 Related work

Point processes have been used extensively for time-to-event data, particularly in survival analysis, where the event in question is typically death (see [1] for an introduction). In survival analysis, censoring, or not observing the process for certain times, is a common factor and more sophisticated tools are used to work around this issue, [3]. Here the Cox proportional hazards with multiplicative factors is used to perform regression on survival data, where an interesting tool, called the partial likelihood is commonplace. I will not use this technique because I am interested in estimating the intensity, and are not particularly interested in the regression setting.

The point process model was extended to recurrent events and the equivalent counting process model is well understood mathematically [2, 13, 23], via filtrations, predictable processes, and martingales. Some early works introduced non-parametric and semi-parametric estimation, [14, 22], and simulation tools, [29], for temporal point processes. Previously, IP traffic has been modelled with point processes, [21]. Event-history data typically takes the form of a point process in time, but events may have a spatial coordinate as well. The general framework of a point process as a random measure over some measurable space easily accommodates space-time processes. Space-time point processes have been extensively analyzed and efficient methods can be found in [17, 32, 35].

Point processes may also be multivariate, which means that there are multiple point processes with possibly interdependent intensities. The lasso has been proposed to estimation the intensity for multivariate point processes [20]. Furthermore, multivariate Hawkes models have seen a surge in interest because multivariate point processes can be used to uncover a latent network of interactions (see [19]). Recent mathematical developments in understanding multivariate Hawkes processes can be found in [5, 15, 41]. The network structure estimation from Hawkes processes has applications to functional connectivity in neuroscience, [31], clustering document streams, [15], high-frequency trading, [6], crime data analysis, [25], and information diffusion, [41]. This setting should not be confused with the point-to-point process, where a network is observed, as opposed to being a latent structure that determines the intensity of a multivariate point process.

Classical network models such as the Erdös-Rényi random graph, see [9], were explored mostly out of mathematical curiosity. It was found that many of the known properties of these preliminary models inadequately modelled real-world phenomena. New network models, such as the preferential attachment model, [8, 27], the small world model, [39], and exponential random graph models [18], were proposed because they reproduced macroscopic properties observed in many real-world networks.

More recently stochastic block models were proposed as a way to incorporate a latent community structure into the network model. [24] introduced the stochastic block model and proposed spectral clustering for recovering block structure. Spectral clustering is based on the idea that the underlying structure of the network can be uncovered by the eigenvectors of certain graph-based matrices (see [37] for a tutorial). Moreover, various clustering algorithms can be used in order to recover latent community structure; one can see various incarnations in [7, 28, 34]. Spectral clustering poses various interesting theoretical questions, most notable, what signal-to-noise ratios are needed to detect community structure with spectrum-based methods. This has been addressed in [4, 11, 16, 26, 36], with a particular emphasis on the information theoretic limits of such problems.

Spatio-temporal networks can have different dynamic components. Vertices may appear and disappear, edges may appear and disappear, and any values or labels may change over time. A time series of graph model was introduced in [38] for detecting anomalies, which is similar to our setting. In another related work to this proposal, latent space models have been proposed for dynamic

networks, [33]. A dynamic stochastic block model was proposed in [40], and an extended Kalman filter was proposed. As mentioned, an incarnation of the point-to-point process model has been introduced in [30], but not in the generality and to the extent considered here. [10] introduced a dynamic network model with continuous time edge events in which the jump times are modelled as exponential random variables but is not analyzed in the context of point processes.

2 METHODOLOGY

The simultaneous estimation of the block structure and the temporal changepoint structure poses several challenges. For massive datasets, storing the tensor $A_{ij}^{(q)}$ in memory is typically impossible, so any optimization algorithms that operate on the full dataset will require distributed implementations. Computational issues aside, it is not at all obvious how one would jointly estimate the changepoints and block model. We propose below a two stage procedure that estimates the community structure from a temporally aggregated network, and then with this block model estimating the temporal changepoints. Furthermore, we will use the Frobenius norm as a approximation for the point-to-point process likelihood. The square error loss has been used as a pseudo-likelihood for point processes in [20]. The point-to-point process assumptions, specifically independent increments, is critical to the proposed methodology. In future work, we plan to explore the use of the true likelihood to form the optimization objectives.

2.1 Community detection

This block model assumption means that there is a low rank structure on $\Lambda^{(q)}$, specifically, that each can be expressed as a linear combination of the matrices,

$$\mathbf{1}_{B_{k_1}^-}\mathbf{1}_{B_{k_2}^+}^T, \quad k_1, k_2 = 1, \ldots, K,$$

where $\mathbf{1}_B$ is the indicator vector over B. This implies that any linear combination of $\Lambda^{(q)}$ can likewise be decomposed as a linear combination of these matrices. Using this fact we will use one specific linear combination, the sum of these matrices, to recover the latent community structure.

The cumulative observed adjacency matrix $A^{(0,T]} := \sum_{q=1}^{M_T} A^{(q)}$, is a matrix of independent Poisson random variables with cumulative intensity $\Lambda^{(0,T]} := \sum_{q=1}^{M_T} \Lambda^{(q)}$. The matrix $A^{(0,T]}$ can be constructed in a distributed fashion, and is a simple database query. We will use adjacency-based spectral clustering on $A^{(0,T]}$, which has the advantage of relying on a singular value decomposition, which has fast iterative implementations. It is not obvious under which conditions this methodology is statistically sub-optimal, and we reserve a theoretical analysis of this method and possible extensions for future work.

With cumulative observed adjacency matrix $A^{(0,T]}$, and a chosen K, we can perform spectral clustering described as follows to detect communities:

(1) Apply singular value decomposition (SVD) on matrix $A^{(0,T]}$ to obtain its best rank-K approximation:

$$A^{(0,T]} \approx U_K D_K V_K^T = \tilde{U}\tilde{V}^T, \text{where } \tilde{U} = U_K D_K^{\frac{1}{2}}, \tilde{V} = V_K D_K^{\frac{1}{2}}.$$

(2) Perform K-means on rows of $\tilde{U}$ and $\tilde{V}$ to get block assignment functions $c(\cdot)$ and $d(\cdot)$:

$$c(i) = k \text{ if } S_i \in B_k^-; \quad d(j) = k \text{ if } S_j \in B_k^+.$$

In this way, separate clusterings of the pickup locations and dropoff locations are discovered. This will be the main vehicle for spatial parameter estimation, so henceforth, we will focus on temporal trends.

2.2 Temporal changepoint estimation

We begin by estimating the base rate $\lambda_0^{(q)}$ as

$$\hat{\lambda}_0^{(q)} = \frac{1}{M_S^2} \sum_{i,j=1}^{M_S} A_{ij}^{(q)}.$$

Throughout the temporal estimation portion, we will divide the counts by this estimated baserate. With the block assignment functions, we can define U and V so that

$$U_{ik} = \begin{cases} \frac{1}{\sqrt{C_k}} & \text{if } S_i \in B_k^-, \text{ i.e. } c(i) = k, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

$$V_{jk} = \begin{cases} \frac{1}{\sqrt{D_k}} & \text{if } S_j \in B_k^+, \text{ i.e. } d(j) = k, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

C_k: the number of sub-areas in the k-th block in B_k^-, $k = 1, 2, ..., K$.
D_k: the number of sub-areas in the k-th block in B_k^+, $k = 1, 2, ..., K$.
Notice that U and V have orthonormal columns, and can be used to reconstruct a low-rank version of each observed adjacency matrix $A^{(q)}$:

$$A^{(q)} \approx \tilde{A}^{(q)} = \hat{\lambda}_0^{(q)} U \Psi^{(q)} V^T, \tag{3}$$

$$\text{i.e. } \tilde{A}_{ij}^{(q)} = \hat{\lambda}_0^{(q)} \frac{\Psi_{k_1,k_2}^{(q)}}{\sqrt{C_{k_1} D_{k_2}}}, \text{ where } c(i) = k_1, d(j) = k_2 \tag{4}$$

it can be shown that the best approximating matrix $\tilde{A}^{(q)}$ has corresponding

$$\hat{\lambda}_0^{(q)} \Psi_{(k_1,k_2)}^{(q)} = \frac{1}{\sqrt{C_{k_1} D_{k_2}}} \sum_{i,j=1}^{M_S} \mathbf{1}_{c(i)=k_1, d(j)=k_2} A_{ij}^{(q)}$$
$$\Leftrightarrow \hat{\lambda}_0^{(q)} \Psi^{(q)} = U^T A^{(q)} V$$

Assuming our estimation of communities U and V are indeed the latent communities, i.e.

$$\Lambda^{(q)} = \lambda_0^{(q)} U \Psi^{(q)} V^\top,$$

our objective function to be minimized could be a group fused LASSO based on the squared error loss on $\{\Psi^{(q)}\}$:

$$F(\{\Psi^{(q)}\}_{q=1}^{M_T}) = \sum_{q=1}^{M_T} \|\Psi^{(q)} - \Phi^{(q)}\|_F^2 + \alpha \sum_{q=1}^{M_T - 1} \|\Phi^{(q+1)} - \Phi^{(q)}\|_F. \tag{5}$$

Here α is a penalty parameter that enforces penalization on change in consecutive $\Phi^{(q)}$'s, and such an objective function will encourage

consecutive $\Phi^{(q)}$'s to be either the same or all elements are different. The larger α is, we expect less change in $\Phi^{(q)}$.

To solve this optimization program, observe that (5) is equivalent to the following problem:

$$\min_{\Phi^{(q)}, Z^{(q)}} \sum_{q=1}^{M_T} \|\Psi^{(q)} - \Phi^{(q)}\|_F^2 + \alpha \sum_{q=1}^{M_T-1} \|Z^{(q)}\|F, \tag{6}$$

$$\text{subject to } Z^{(q)} = \Phi^{(q+1)} - \Phi^{(q)} \tag{7}$$

by introducing a new variable, $Z^{(q)}$, to capture the difference between successive Φ's.

With this formulation we will apply the Alternating Direction Method of Multipliers (ADMM) procedure to solve the problem. The advantages of this procedure are that ADMM and other augmented Lagrangian based algorithm makes the transformed primal more well-conditioned thus yields better convergence conditions [12], and the updates for the variables can be carried out in a parallel fashion.

We can look at the solution $Z^{(q)}$ given by the procedure to determine if t_{q+1} is a changepoint. With these changepoints, we re-estimate the demand matrix to remove additional bias due to the fusion penalty. We use

$$\hat{\Lambda}^{(q)} = \frac{\hat{\lambda}_0^{(q)}}{q_2 - q_1} \sum_{q'=q_1}^{q_2} UU^\top A^{(q')} VV^\top,$$

where τ_{q_1}, τ_{q_2} are two consecutive changepoints and t_q is contained in between.

2.3 Model selection

In our method, there are mainly two tuning parameters that need to be selected, K, number of communities in community detection, and α in the penalty to the change in intensities. We can either choose these parameters under certain rough guidelines, or by tuning on a validation set.

To choose K, we can look at variance explained in the observed counting matrix $A^{(0,T]}$ for different values of K and choose a reasonable one that satisfies the compression budget. Similarly, to choose α, we can estimate the changepoints over a range of candidate values of α and choose the one that gives the appropriate amount of changepoints fitting the compression budgets.

Another method is to use a validation set, by selecting a portion of trips to form a validation set while the remaining forms the training set. With several candidate values of K and α, we can run the analysis on the training set to get changepoints and estimated intensities $\hat{\Lambda}^{(q)}$. Then squared error loss can be used to select a best combination of K and α. Because we can decompose the square error loss into a component due to the community detection and the changepoint estimation, we can use the following:

$$\sum_{q=1}^{M_T} ||\Psi_{\text{test}}^{(q)} - \Phi^{(q)}||_F^2 \tag{8}$$

to select the number of changepoints.

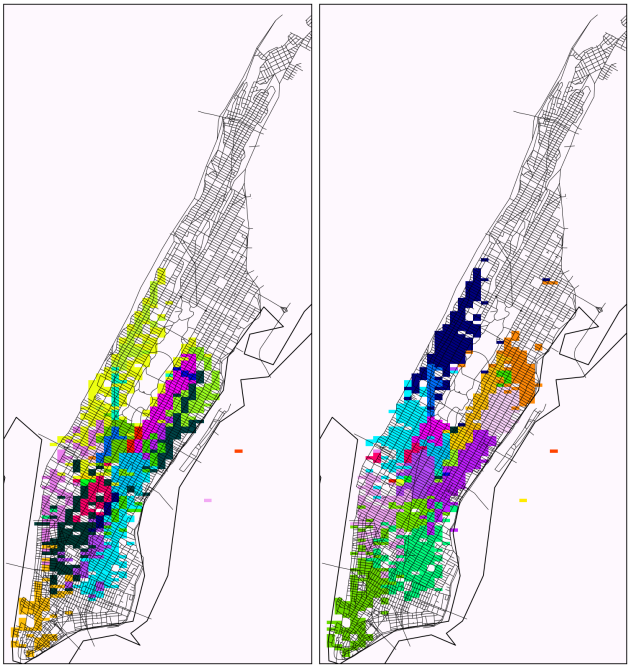

Figure 4: Departure blocks (left figure) and arrival blocks (right figure) estimated with number of blocks, $K = 30$.

3 EXPERIMENTAL RESULTS

To demonstrate the methods described in Section 2, we use the yellow taxi trip records during January, 2015, consisting of over ten million trips. The trips are randomly divided equally to form the training and test datasets.

The choice of fine mesh used in this section to aggregate the trips spatially is to partition the region shown in Figure 2 into The fine mesh that was used to aggregate trips as a preprocessing step was composed of $M_S = 200 \times 40$ grid regions, making each sub-area roughly the size of a city block (you can see these in Figure 2). The temporal resolution was such that the time of a week is divided into $M_T = 24 \times 7$ equal-length time intervals with $\Delta t = 12$ minutes.

The result of K-means for $K = 30$ is depicted in Figure 4, where different colors indicate different cluster membership. Recall that separate cluster models are fitted for drop-off locations and pick-up locations. While the cluster memberships for departures and arrivals have similarities, they are not identical, indicating that some regions may have different roles during a pick-up and a drop-off. The detected communities are consistent with the neighborhood structure of Manhattan. Separate communities can be seen for Times Square, Broadway Ave. from Columbus circle to the Lincoln Center, the east side of Central Park where there are many museums, the lower east side, West Greenwich village, etc. In fact, the community structure seems to be more dependent on neighborhood boundaries and not merely on geographic location.

Here we explore a set of tuning parameters: $K = 30$ or 50, α takes one of 21 equidistant values from 0 to 100. With a combination of K and α, suppose there are p changepoints detected, then the number of parameters to be stored would be $p \cdot K^2$, apart from storing

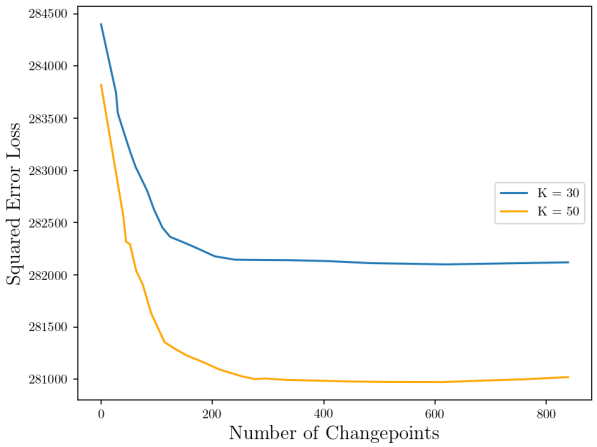

Figure 5: Squared error loss of the test set plotted against number of changepoints to store after detecting changepoints. $K = 30$ (blue curve) and $K = 50$ (orange curve) are considered.

the location of these changepoints. Figure 5 shows the number of changepoints against squared error loss of the test set, $\sum_{q=1}^{M_T} ||A^{(q)}_{\text{train}} - \hat{\Lambda}^{(q)}||_F^2$. It shows that with a relatively large compression budget or transmission rate, the loss can be reduced with increased number of changepoints. This effect tails off or even is reversed when number of changepoints reaches a fairly large amount, which is due to overfitting. With such a plot, users of our method are able to select tuning parameters by balancing their compression needs and the amount of affordable loss of accuracy.

4 DISCUSSION

Demand estimation and compression for large spatio-temporal networks poses many theoretical and practical challenges. We have introduced a novel framework called the point-to-point process which allows us to use point processes to model spatio-temporal networks. By estimating a spatial block model and temporal changepoint model, we obtain a compressed estimate of the point-to-point process intensity. For mobile applications this feature is of paramount importance, and we show that this is possible with a substantial compression ratio.

Applications of point-to-point process models are ubiquitous in modern transactional databases. This framework can be applied to internet packet data, financial transaction records, and other transportation networks. Several extensions of this work are possible, particularly using the likelihood instead of the square error loss in this methodology. The assumptions made in this work, the independent increments and constant spatial community structure, are specific to this application. Other applications will motivate various extensions, such as modelling temporal dependence, changing community structure, and trend-filtering of temporal structure.

REFERENCES

[1] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. 2008. *Survival and event history analysis: a process point of view.* Springer Science & Business Media.
[2] Per Kragh Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. 2012. *Statistical models based on counting processes.* Springer Science & Business Media.
[3] Per Kragh Andersen and Richard David Gill. 1982. Cox's regression model for counting processes: a large sample study. *The annals of statistics* (1982), 1100–1120.
[4] Ery Arias-Castro, Nicolas Verzelen, and others. 2014. Community detection in dense random networks. *The Annals of Statistics* 42, 3 (2014), 940–969.
[5] Emmanuel Bacry, Stéphane Gaïffas, and Jean-François Muzy. 2015. A generalization error bound for sparse and low-rank multivariate Hawkes processes. *arXiv preprint arXiv:1501.00725* (2015).
[6] Emmanuel Bacry and Jean-François Muzy. 2014. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance* 14, 7 (2014), 1147–1166.
[7] Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. 2011. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*. 954–962.
[8] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.
[9] Béla Bollobás. 1998. Random graphs. In *Modern Graph Theory*. Springer, 215–252.
[10] Carter T Butts. 2008. A relational event framework for social action. *Sociological Methodology* 38, 1 (2008), 155–200.
[11] Cristina Butucea, Yuri I Ingster, and others. 2013. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli* 19, 5B (2013), 2652–2688.
[12] Antonin Chambolle and Thomas Pock. 2011. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision* 40, 1 (2011), 120–145.
[13] Daryl J Daley and David Vere-Jones. 2007. *An introduction to the theory of point processes: volume II: general theory and structure.* Springer Science & Business Media.
[14] Peter Diggle. 1985. A kernel method for smoothing point process data. *Applied statistics* (1985), 138–147.
[15] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J Smola, and Le Song. 2015. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 219–228.
[16] Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. 2013. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* 34, 1 (2013), 23–39.
[17] P Fishman and D Snyder. 1976. The statistical analysis of space-time point processes. *IEEE Transactions on Information Theory* 22, 3 (1976), 257–274.
[18] Ove Frank and David Strauss. 1986. Markov graphs. *Journal of the american Statistical association* 81, 395 (1986), 832–842.
[19] Eric Hall and Rebecca Willett. 2012. Tracking Dynamic Point Processes on Networks. *IEEE Transactions on Information Theory* 62 (2012). Issue 7.
[20] Niels Richard Hansen, Patricia Reynaud-Bouret, Vincent Rivoirard, and others. 2015. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli* 21, 1 (2015), 83–143.
[21] Alexander Klemm, Christoph Lindemann, and Marco Lohmann. 2003. Modeling IP traffic using the batch Markovian arrival process. *Performance Evaluation* 54, 2 (2003), 149–173.
[22] DY Lin, LJ Wei, I Yang, and Z Ying. 2000. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62, 4 (2000), 711–730.
[23] JA McFadden. 1965. The entropy of a point process. *J. Soc. Indust. Appl. Math.* 13, 4 (1965), 988–994.
[24] Frank McSherry. 2001. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on.* IEEE, 529–537.
[25] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. 2012. Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* (2012).
[26] Raj Rao Nadakuditi and Mark EJ Newman. 2012. Graph spectra and the detectability of community structure in networks. *Physical review letters* 108, 18 (2012), 188701.
[27] Mark EJ Newman. 2001. Clustering and preferential attachment in growing networks. *Physical review E* 64, 2 (2001), 025102.
[28] Andrew Y Ng, Michael I Jordan, Yair Weiss, and others. 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* 2 (2002), 849–856.
[29] Yosihiko Ogata. 1981. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory* 27, 1 (1981), 23–31.
[30] Patrick O Perry and Patrick J Wolfe. 2013. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75, 5 (2013), 821–849.

[31] Patricia Reynaud-Bouret, Vincent Rivoirard, and Christine Tuleau-Malot. 2013. Inference of functional connectivity in neurosciences via Hawkes processes. In *1st IEEE Global Conference on Signal and Information Processing*.
[32] Frederic Paik Schoenberg. 2015. A note on the consistent estimation of spatial-temporal point process parameters. *Statistica Sinica* (2015).
[33] Daniel K Sewell and Yuguo Chen. 2015. Latent space models for dynamic networks. *J. Amer. Statist. Assoc.* 110, 512 (2015), 1646–1657.
[34] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22, 8 (2000), 888–905.
[35] David Vere-Jones. 2009. Some models and procedures for space-time point processes. *Environmental and Ecological Statistics* 16, 2 (2009), 173–195.
[36] Nicolas Verzelen, Ery Arias-Castro, and others. 2015. Community detection in sparse random networks. *The Annals of Applied Probability* 25, 6 (2015), 3465–3510.
[37] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing* 17, 4 (2007), 395–416.
[38] Heng Wang, Minh Tang, Youngser Park, and Carey E Priebe. 2014. Locality statistics for anomaly detection in time series of graphs. *IEEE Transactions on Signal Processing* 62, 3 (2014), 703–717.
[39] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of small-world networks. *nature* 393, 6684 (1998), 440–442.
[40] Kevin S Xu and Alfred O Hero. 2014. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing* 8, 4 (2014), 552–562.
[41] Ke Zhou, Hongyuan Zha, and Le Song. 2013. Learning Social Infectivity in Sparse Low-rank Networks Using Multi-dimensional Hawkes Processes.. In *AISTATS*, Vol. 13. 641–649.