Universal Models of Multivariate Temporal Point Processes (With Supplementary Appendix Containing Proofs)

Asela Gunawardana
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
aselag@microsoft.com

Christopher Meek
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
meek@micrsoft.com

Abstract

With the rapidly increasing availability of event stream data there is growing interest in multivariate temporal point process models to capture both qualitative and quantitative features of this type of data. Recent research on multivariate point processes have focused in inference and estimation problems for restricted classes of models such as continuous time Bayesian networks, Markov jump processes, Gaussian Cox processes, and Hawkes Processes.

In this paper, we study the expressive power and learnability of Graphical Event Models (GEMs) — the analogue of directed graphical models for multivariate temporal point processes. In particular, we describe a set of Graphical Event Models (GEMs) and show that this class can universally approximate any smooth multivariate temporal point process. We also describe a universal learning algorithm for this class of GEMs and show, under a mild set of assumptions, learnability results for both the dependency structures and distributions in this class. Our consistency results demonstrate the possibility of learning about both qualitative and quantitative dependencies from rich event stream data.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

1 Introduction

There has been an explosion in the availability of event stream data which has been collected to explore the dynamics of a wide variety of systems behavior including social networks (Du et al., 2013), biochemical networks (Golightly and Wilkinson, 2006), patient health (Weiss and Page, 2013), and computers in datacenters (Gunawardana et al., 2011). A wide variety of alternative statistical models have been advanced and applied to these data including continuous time Bayesian networks (Nodelman et al., 2002), graphical event models (e.g. Gunawardana et al., 2011; Weiss and Page, 2013), Markov jump processes (Rao and Teh, 2013), Gaussian Cox processes (e.g. Adams et al., 2009; Lian et al., 2015), and Hawkes Processes (e.g. Zhou et al., 2013). Most work on modelling such data address questions about choosing representations, learning algorithms, and inference techniques for particular model classes, as well as how well various model classes capture the dynamics of data from systems.

In contrast, we address the question of whether there exist classes of models that can approximate any temporal marked point process, and if so, if such models are learnable. For other learning tasks such as classification, regression, and modeling discrete-alphabet sequences, there are results on universal approximation, and on universal consistency or learnability that show that certain model classes (such as k-nearest neighbor, support vector machines with universal kernels, and context trees for variable length Markov chains) can approximate any distribution of interest and can estimate such a distribution with arbitrary accuracy from sufficient data. In this paper, we give such results for temporal marked point processes using a class of graphical event models. Similar to a directed graphical model, graphical event models capture both qualitative structural aspects of the temporal dependencies as well as quantitative aspects of system dynamics.

Our main contributions are theoretical results about a class of graphical event models that we call Recursive Timescale Graphical Event Models. In particular, we prove a universal approximation theorem that shows that Recursive Timescale Graphical Event Models can approximate any sufficiently smooth finite-horizon marked point process. We then provide a universal learning algorithm for recursive timescale graphical event models and give asymptotic consistency results for this algorithm. These results show that such models can, in principle, be learned from sufficient data. In particular, we give structural consistency results that show that the dependency structure of such processes are correctly learned, parametric consistency results that show that the marginal distributions (and thus the process itself) is correctly learned, and predictive consistency results that show that the learned model makes correct predictions about the future. Our consistency results demonstrate the possibility of learning about both qualitative and quantitative dependencies from rich event stream data.

2 Related Work

The study of universal learners goes back at least to Cover and Hart (1967) and Stone (1977) who studied the universality of k-nearest neighbor for classification and regression respectively. Hornik (1989) gave a universal approximation result for neural networks, but to our knowledge, there are few consistency results for learning them. In particular, there remains a need for theoretical guidance for selecting the topology and number of hidden units. More recently Steinwart (2001) characterized universal kernels and studied the consistency of SVMs using such kernels.

In the case of discrete symbol sequence modelling, the results of Weinberger et al. (1995), Bühlmann and Wyner (1999) and Csiszár and Talata (2006) establish that variable length Markov models are universal models for arbitrary ergodic discrete symbol sources, and that the context tree algorithm can consitently learn them. This work extends previous work on universal source coding (Rissanen and Langdon, 1981; Rissanen, 1983).

To our knowledge, the results we present here are the first result on a universal approximating class for temporal marked point processes, and the first learnability result for such a class.

Finally, although we do not address inference in this paper, we note that the inference algorithms of Qin and Shelton (2015) can be applied to the models yielded by the learning algorithms we present here.

3 Background

Our data consists of a stream of events $(t, l) \in \mathbb{R}^+ \times \mathcal{L}$, each of which has a timestamp $t > 0$ and a label l taken from a finite label vocabulary $\mathcal{L}$. Thus the data is a sequence $(t_1, l_1), \cdots, (t_i, l_i), \cdots$ with strictly increasing times. We write x_{t^*} for the sequence of events $\{(t_i, l_i) : t_i < t^*\}$ until time t^*, and wish to model $p(x_{t^*}|t^*)$ for any given t^*. A family of distributions $\{p(x_{t^*}|t^*)\}_{t^*>0}$ is consistent with each other if their marginals agree, and such a family is known as a *marked point process* (m.p.p.) P (Daley and Vere-Jones, 2003).

3.1 Graphical Event Models

A *Graphical Event Model* (GEM) $\mathcal{G}$ is a directed graph $\mathcal{G} = (\mathcal{L}, E)$. A GEM $\mathcal{G}$ defines a family of m.p.p.s whose likelihood of the data x_{t^*} can be written as

$$p(x_{t^*}|t^*) = \prod_{i=1}^{|x_{t^*}|} \lambda_{l_i}(t_i|h_i) \prod_{i=1}^{|x_{t^*}|+1} e^{-\sum_{l\in\mathcal{L}} \int_{t_{i-1}}^{t_i} \lambda_l(\tau|h_i)d\tau}$$

where we use h_i to denote the ith history $h_i = (t_1, l_1), \cdots, (t_{i-1}, l_{i-1})$, and have used the conventions $t_0 = 0, t_{n+1} = t^*$, and where $\lambda_l(t|h) > 0$ is the *conditional intensity* of the label l at time t given the history h which governs how the occurrence of events with label l at time t depends on the history h. This representation is quite general, and the conditional intensity function can be written in terms of conditional densities and conditional probabilities as:

$$\lambda_l(t|h_i) = \frac{p(l_i = l, t_i = t|h_i)}{P(t_i > t|h_i)}$$

For example, PCIMs (Gunawardana et al., 2011), CPCIMs (Parikh et al., 2012), CTBNs (Nodelman et al., 2002) and MFPPs (Weiss and Page, 2013) can all be represented as GEMs.

The m.p.p.s defined by the GEM $\mathcal{G}$ have the following property:

Definition 1. A m.p.p. P is *Markov* with respect to a GEM $\mathcal{G}$ if its conditional intensity functions $\lambda_l(t|h)$ satisfy

$$\lambda_l(t|h) = \lambda_l(t|[h]_{\mathrm{Pa}(l)})$$

where $\mathrm{Pa}(l)$ are the parents of l in $\mathcal{G}$, and $[h]_{\mathcal{K}} = \{(t, l) \in h : l \in \mathcal{K}\}$ is the subset of events in h whose labels are in $\mathcal{K}$. This is analogous to Bayesian Networks where the conditional probability of a variable given the preceding variables depends only on its parents.

Just as any joint distribution over a set of variables can be specified by a fully connected Bayesian network, any m.p.p. with labels $\mathcal{L}$ can be represented by

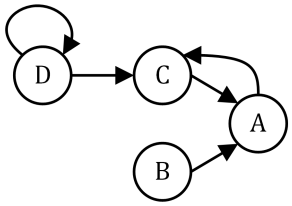

Figure 1: The GEM of example 1

a fully connected GEM, under mild regularity conditions (Daley and Vere-Jones, 2003).

Example 1. Consider the GEM illustrated in Figure 1. `A` denotes the alarm triggering. The alarm can be triggered by either a break-in (event `B`) or by the cat jumping onto the window sill (event `C`). The cat typically hides after she has been taken to see the vet, and is thus less likely to jump on the window sill. The event `D` is the event of the cat being taken to see the vet, or from her point of view, a visit to the doctor. Dependencies between events can be read off the GEM. In this example, break-ins `B` do not depend on any event in the history, including past break-ins. The rate of break-ins $\lambda_{\mathtt{B}}(t|h)$ is therefore constant, and break-ins form a homogeneous Poisson process. The rate $\lambda_{\mathtt{D}}(t|h)$ of visits to the doctor `D` depends on past visits, and reflects the self-excitatory and self-inhibitory effects due to the need or lack thereof for follow-up care. The cat's rate $\lambda_{\mathtt{C}}(t|h)$ of jumping on the window sill depends only on the past only through alarms `A` and visits to the doctor `D`.

4 Universal Approximation

In order to show our approximation result, we first define a notion of a Timescale GEM, which is a GEM where the temporal range and granularity of each dependency is made explicit. We then define a nested family of such Timescale GEMs that allow longer and longer range dependencies with finer and finer temporal granularity. We then show that a member of this family with sufficiently long-range dependencies with sufficiently fine temporal granularity will suffice to approximate any m.p.p. meeting some mild assumptions.

4.1 Timescale Graphical Event Models

A *Timescale GEM* (TGEM) augments each edge of a GEM with a *timescale* which specifies the finite temporal horizon and the granularity of the dependency represented by that edge. Formally, a *timescale* is a set T of half-open intervals $I \subset \mathbb{R}^+$ of the form $(a, b]$ that form a partition of some interval $(0, t_h]$. We call the highest value contained in $\cup_{I \in T} I$ the *horizon* of T denoted $t_h(T)$.

A TGEM $M = (\mathcal{G}, \mathcal{T})$ consists of a graph $\mathcal{G} = (\mathcal{L}, E)$ over $\mathcal{L}$ and a set of timescales $\mathcal{T} = \{T_e\}_{e \in E}$ corresponding to the edges E of the graph $\mathcal{G}$. The TGEM M specifies m.p.p.s with conditional intensity functions that have parameters

$$\lambda_l(t|h) = \lambda_{l,c_l(h,t)}$$

indexed by *parent count vectors* $c_l(h,t)$ of thresholded counts over the intervals in the timescales of the parents of l

$$c_l(h,t) = [c_{l',I}(h,t)]_{l' \in \mathrm{Pa}(l), I \in T_{(l',l)}}$$

with elements

$$c_{l',I}(h,t) = \left\lfloor \sum_{(t',l') \in h} \mathbf{1}_I(t - t') \right\rfloor_k$$

which count the number of events with parent label l' in the history h within the timescale interval I of time t, truncated at some given threshold. This is analogous to conditional probability parameters in Bayesian networks being indexed by parent states. We use C_l to denote the set of all possible parent count vectors of label l. For this paper we assume that all TGEMs have a threshold of 1 but note that all our results apply to other choices of threshold.

Thus, a TGEM defines a set of m.p.p.s as follows:

Definition 2. A m.p.p. P is *Markov* with respect to a TGEM $M = (\mathcal{G}, \mathcal{T})$ (written $P \in \mathrm{Markov}(M)$) if $P \in \mathrm{Markov}(\mathcal{G})$ and if there exist non-negative parameters $\{\lambda_{l,j}\}_{l \in \mathcal{L}, j \in C_l}$ such that $\lambda_l(t|h) = \lambda_{l,c_l(h,t)}$.

For $P \in \mathrm{Markov}(M)$ the likelihood simplifies to

$$p(x_{t^*}|t^*) = \prod_{l \in \mathcal{L}} \prod_{j \in C_l} \lambda_{l,j}^{n_{t^*,l,j}(x_{t^*})} e^{-\lambda_{l,j} d_{t^*,l,j}(x_{t^*})}$$

where the sufficient statistics $n_{t^*,l,j}(x_{t^*})$ and $d_{t^*,l,j}(x_{t^*})$ are the count of l-events and the duration, respectively, when the parent count vector was equal to j, and are given by

$$n_{t^*,l,j}(x_{t^*}) = \sum_{i=1}^{|x_{t^*}|} \mathbf{1}(l_i = l)\mathbf{1}(c_l(h_i, t_i) = j)$$

$$d_{t^*,l,j}(x_{t^*}) = \sum_{i=1}^{|x_{t^*}|+1} \int_{t_{i-1}}^{t_i} \mathbf{1}(c_l(h_i, \tau) = j) d\tau$$

For our consistency results, we assume that our data is generated from a TGEM with strictly positive parameters. This is required for Lemma 3 and results that depend on it. This assumption ensures that the observed dynamics of the process will repeat.

Example 2. We extend the GEM of Example 1 to a TGEM by specifying timescales for edges in Figure 1. The cat hides for an hour after a visit D to the doctor, and for 5 or 10 minutes after hearing the alarm. Thus, the timescales of the corresponding edges are $T_{\mathtt{A}\to\mathtt{C}} = \{(0,5],(5,10]\}$ and $T_{\mathtt{D}\to\mathtt{C}} = \{(0,60]\}$. The label C has a three-dimensional binary parent count vector $c_{\mathtt{C}}(h,t)$ that encodes whether there was an alarm A in the intervals $[t-5,t)$ and $[t-10,t-5)$, and whether there was a doctor's visit D in the interval $[t-60,t)$. In this case, $C_{\mathtt{C}} = \{0,1\}^3$ and if $c_{\mathtt{C}}(h,t) = j$ with $j = [0,1,1]^T$, there was no alarm in $[t-5,t)$, there was an alarm in $[t-10,t-5)$, and there was a doctor's visit in $[t-60,t)$.

4.2 Recursive Timescale GEMs

Our goal is to learn the edges and the appropriate timescale for each edge in a TGEM. We will do so by using a set of operators for refining TGEMs defined below:

Definition 3. The *add edge operator* $O_{\text{add},e}$ adds the edge e to a model. It is formally defined for models $M = (\mathcal{G},\mathcal{T})$ such that $e \notin E_{\mathcal{G}}$ by $O_{\text{add},e}(M) = (\mathcal{G}',\mathcal{T}')$ with $E_{\mathcal{G}'} = E_{\mathcal{G}} \cup \{e\}$, $T'_e = T_0$, and $T'_{e'} = T_{e'}$ for $e' \in E_{\mathcal{G}}$, where $T_{e'}$ and $T'_{e'}$ are the timescales of edge e' in models M and $\mathcal{M}'$ respectively. We use $T_0 = \{(0,c]\}$ for some constant c.

Definition 4. The *split operator* $O_{\text{split},e,(a,b]}$ splits the interval $(a,b]$ in the timescale of edge e. It is formally defined for models $M = (\mathcal{G},\mathcal{T})$ such that $e \in E_{\mathcal{G}}$ and $(a,b] \in T_e \in \mathcal{T}$ by $O_{\text{split},e,(a,b]}(M) = (\mathcal{G},\mathcal{T}')$ with $T'_e = \left\{\left(a,\frac{a+b}{2}\right],\left(\frac{a+b}{2},b\right]\right\} \cup T_e \setminus (a,b]$.

Definition 5. The *extend operator* $O_{\text{extend},e}$ extends the horizon of edge e. Formally, for models $M = (\mathcal{G},\mathcal{T})$ such that $e \in E_{\mathcal{G}}$ and $t_h(T_e) = t_h$, it gives $O_{\text{extend},e}(M) = (\mathcal{G},\mathcal{T}')$ with $T'_e = T_e \cup \{(t_h, 2t_h]\}$.

The family of *Recursive Timescale GEMs* (RTGEMs) is defined recursively to be the finite closure of the empty model $M_0 = ((\mathcal{L},\{\}),\{\})$ under the add edge, split, and extend operators. That is, RTGEMs include all GEMs reachable from M_0 via any finite sequence of refinement operations.

RTGEMs are a rich class of models that can approximate arbitrary non-explosive non-deterministic smooth m.p.p.s with finite horizons:

Theorem 1. *Suppose P is a stationary m.p.p. whose conditional intensity functions $\lambda_l(t|h)$ are bounded above and bounded away from zero, are Lipschitz continuous in t and h (i.e. in t and in the times of the events in h), and that there exits $t_h \geq 0$ such that $\lambda_l(t|h)$ does not depend on events in h at times earlier than $t - t_h$. Then, for any $\epsilon > 0$ there exists an RTGEM $\hat{M} \in \mathcal{M}$ and $\hat{P} \in Markov(\hat{M})$ such that $\frac{1}{t^*} D(P_{t^*} || \hat{P}_{t^*}) < \epsilon$.*

Note that a stationary m.p.p. is one where the conditionaly intensity functions are invarient under translation: $\lambda_l\left(t|(t_1,l_1),\cdots,(t_i,l_i)\right) = \lambda_l\left(t-\tau|(t_1-\tau,l_1),\cdots,(t_i-\tau,l_i)\right)$ for all τ. Also, note that a non-explosive m.p.p. is one where $\lim_{i\to\infty} t_i = \infty$ a.s. In contrast, explosive m.p.p.s are ones where the inter-event times can shrink fast enough to allow an inifinite number of events in a finite duration.

5 Consistency

We now give asymptotic consistency results for RTGEMs. First we give a parametric and predictive consistency result for TGEMs. That is, we show that we can learn the correct parameters if the dependency structure and timescales are known. We then go on to give a structural consistency result that the correct TGEM, i.e. the correct dependency structure and timescales can be learned. In order to do so, we first show some results about the structure of RTGEMs that are needed to formalize what is meant by the "correct" TGEM.

5.1 Parametric Consistency

Given a model M, the maximum likelihood estimate (m.l.e.) for M is

$$\hat{\lambda}_{t^*,l,j}(x_{t^*}) = \frac{n_{t^*,l,j}(x_{t^*})}{d_{t^*,l,j}(x_{t^*})}$$

for $l \in \mathcal{L}, j \in C_l$. The m.l.e. is consistent, and leads to consistent predictions of the future:

Theorem 2. *Suppose $P \in Markov(M)$. Then, $\hat{\lambda}(x_{t^*}) \to \lambda$ in probability, and for any $\epsilon > 0, \Delta > 0$,*

$$P\left(\left|\log\frac{p(x_{t^*+\Delta}|x_{t^*},M)}{\hat{p}(x_{t^*+\Delta}|x_{t^*},M)}\right| > \epsilon\right) \to 0$$

as $t^* \to \infty$.

This results from the direct application of two concentration lemmas. The first gives that the duration sufficient statistics per unit time converge to their long run expected duration rates:

Lemma 3. *For any P such that $\exists M \in \mathcal{M} : P \in Markov(\mathcal{M})$, and every $l \in \mathcal{L}, j \in C_l$, there exists $r_{l,j} > 0$ such that for every $\epsilon > 0, \delta > 0$ there exists $t > 0$ s.t. $P\left(\left|\frac{d_{t^*,l,j}(x)}{t^*} - r_{l,j}\right| < \delta\right) > 1 - \epsilon$ for $t^* > t$.*

The second gives that given the durations, the ratios of the sufficient statistics concentrate around the corresponding parameters:

Lemma 4. *For any m.p.p. P such that $\exists M \in \mathcal{M}$: $P \in Markov(M)$ and any $l \in \mathcal{L}, j \in C_l$, there exists a constant $K > 0$ s.t. for all $\epsilon > 0$,*

$$P\left(\left|\frac{n_{t^*,l,j}(x)}{d_{t^*,l,j}(x)} - \lambda_{l,j}\right| > \epsilon \middle| d_{t^*,l,j}(x) = \hat{d}_{t^*,l,j}\right) < 2\left(1 - \Phi\left(\epsilon\sqrt{\frac{\hat{d}_{t^*,l,j}}{\lambda_{l,j}}}\right)\right) + \frac{K}{\lambda_{l,j}\hat{d}_{t^*,l,j}}$$

where $\Phi(\cdot)$ is the standard normal CDF.

5.2 The Structure of RTGEMs

In order to formally state our results on structural consistency, we first show the existence of a minimal TGEM that can represent a m.p.p. P. Structural consistency will then be defined in terms of finding this minimal TGEM. In this section, we introduce the notions of one TGEM being a refinement of another, which we will use when searching for the minimal TGEM.

It can be shown that each $M \in \mathcal{M}$ is characterized by a set $\mathcal{O}_M$ of operators that take M_0 to M, and that every path from M_0 to M uses each of the operators in $\mathcal{O}_M$. For example, for a particular interval to be present in an TGEM, the relevant edge must first be added, and the horizon extended until it covers the interval in question. The timescales then need to be refined until the interval is obtained. All these operations are essential for arriving at a TGEM containing the interval in question. It can also be shown that there is a unique *minimal model* $M^*(P)$ that represents each P that can be represented in $\mathcal{M}$:

Proposition 5. *Let P be such that $\exists M \in \mathcal{M}$ for which $P \in Markov(M)$. Then, there exists a unique minimal $M^*(P) \in \mathcal{M}$ such that $P \in Markov(M^*(P))$ and $M' \geq M^*(P)$ for all M' for which $P \in Markov(M')$.*

RTGEMs have a rich recursive structure that we make use of in the rest of the paper. We say that $M \in \mathcal{M}$ is an *immediate refinement* of $M' \in \mathcal{M}$ (denoted $M \succ M'$) if there exists an add edge, split, or extend operator O such that $M = O(M')$. $M \in \mathcal{M}$ is a *refinement* of $M' \in \mathcal{M}$ (denoted $M > M'$) if there exists a sequence of immediate refinements that take M' to M. Conversely, we say that M' is a *projection* of M. Thus, by definition, $M \geq M_0$ for every $M \in \mathcal{M}$.

The parent count vectors that index the parameters of RTGEMs are also recursively related. For $M > M'$, the parent count vector $c'_l(h,t)$ defined for each $l \in \mathcal{L}$ by model M' can be obtained by summing and truncating the elements of the parent count vector $c_l(h,t)$ defined by M. That is, $c'_l(h,t)$ is a nonlinear projection of $c_l(h,t)$. We denote this by $c'_l(h,t) = \pi^M_{M'}(c_l(h,t))$ and when it is clear from context which models we are projecting from and to, we write $c'_l(h,t) = \pi(c_l(h,t))$.

The recursive projection relationship between parent count vectors gives us that the sufficient statistics are also computable as recursive projections:

Proposition 6. *Let $M, M' \in \mathcal{M} : M > M'$, with parent count vector sets C_l and C'_l respectively. Then, for all $l \in \mathcal{L}, j' \in C'_l$,*

$$\begin{aligned} n_{t^*,l,j'}(x_{t^*}) &= \sum_{j \in \pi^{-1}(j')} n_{t^*,l,j}(x_{t^*}) \\ d_{t^*,l,j'}(x_{t^*}) &= \sum_{j \in \pi^{-1}(j')} d_{t^*,l,j}(x_{t^*}) \end{aligned} \tag{1}$$

5.3 Structural Consistency

We take a Bayesian Information Criterion (BIC) approach to structure learning.

We define the BIC score of an RTGEM M as

$$\mathcal{S}_{t^*}(M) = \log p(x_{t^*}|t^*; M, \hat{\lambda}_{t^*}(x_{t^*})) - \sum_{l \in \mathcal{L}} |C_l| \log t^*$$

The following model selection theorem implies that model selection under the BIC score above is asymptotically consistent for the minimal model $M^*(P)$ of a process P:

Theorem 7. *Suppose $M \neq M' \in \mathcal{M}$ with $P \in Markov(M)$ and with either $P \notin Markov(M')$ or $M' > M$. Then,*

$$\lim_{t^* \to \infty} P\left(\mathcal{S}_{t^*}(M) > \mathcal{S}_{t^*}(M')\right) = 1$$

In particular, this holds for any $M' \neq M$ when $M = M^*(P)$.

The consistency result above directly gives that if $M \in \mathcal{M}$ is such that $P \in \text{Markov}(M)$, starting at M and greedily removing distinctions by applying the function BackwardSearch() of Algorithm 1 yields the minimal model that can represent P:

Corollary 8. *Suppose that $M \in \mathcal{M}$ and let P be such that $M = M^*(P)$. Let $\bar{M} \in \mathcal{M} : P \in Markov(\bar{M})$, and let $\hat{M}_{t^*}$ be the result of applying* BackwardSearch() *to $\bar{M}$ at time t^*. Then, $P(\hat{M}_{t^*} = M) \to 1$ as $t^* \to \infty$.*

This is analogous to results on Backward Equivalence Search (BES) for Bayesian networks (Chickering and Meek, 2002). However, in that case, it is possible to

Algorithm 1 BACKWARDSEARCH(M)

```
repeat
  coarsened ← false
  M' ← {M' ∈ M : M ≻ M'}
  for all M' ∈ M' do
    if S(M') − S(M) > 0 then
      M ← M',   coarsened ← true
      break
    end if
  end for
until not coarsened
return M
```

initialize BES with the fully connected network, which is guaranteed to contain any joint distribution. In contrast, the family of RTGEMs is infinite, and there is no maximal model that can represent any m.p.p., making this result less useful.

In analogy to Forward Equivalence Search (FES) for Bayesian networks (Chickering and Meek, 2002), we can attempt to find a model M such that $P \in \text{Markov}(M)$ by starting with the empty model M_0 and refining it to greedily improve the BIC score $\mathcal{S}(\cdot)$. However, this procedure does not succeed for all m.p.p.s P that have a representation within $\mathcal{M}$. This is because the dynamics of the process can balance out and obscure distinctions in the process P, so that greedy refinement fails. This is illustrated in the following example:

Example 3. Let M be specified as $\mathcal{L} = \{\text{A}, \text{B}\}, E = \{\text{A} \to \text{B}\}, T_{\text{A}\to\text{B}} = \{(0,1],(1,2]\}$. Every $P \in \text{Markov}(M)$ is parametrized by a single parameter for the conditional intensity of events with label A, which form a homogeneous Poisson process, and 4 parameters for the conditional intensity of events with label B, for the cases where there are 0 or at least 1 event with label A in each of the intervals $t-(0,1]$ and $t-(1,2]$. Let such a P be specified with parameters $\lambda_{\text{A}} = \log 4, \lambda_{\text{B},00} = 6, \lambda_{\text{B},01} = 2, \lambda_{\text{B},10} = 9$, and $\lambda_{\text{B},11} = 1$. A successful greedy search would start with the initial model M_0 with $E_0 = \{\}$, refine it to M_1 by adding the edge $\text{A} \to \text{B}$ with timescale $T_{\text{A}\to\text{B}} = T_0 = \{(0,1]\}$, and then extend the timescale to obtain M. M_0 has two parameters λ_{A} and λ_{B} while M_1 has three parameters λ_{A}, $\lambda_{\text{B},0}$, and $\lambda_{\text{B},1}$. We show that the dynamics of the process P obscure the benefit of adding the edge $\text{A} \to \text{B}$ with timescale $T_{\text{A}\to\text{B}} = \{(0,1]\}$ to M_0 and distinguishing between $\lambda_{\text{B},0}$ and $\lambda_{\text{B},1}$.

From Lemmas 3 and 4 and from Proposition 6, we can show that

$$\hat{\lambda}_{t^*,\text{B},0}(x) \to \frac{\lambda_{\text{B},00} r_{\text{B},00} + \lambda_{\text{B},01} r_{\text{B},01}}{r_{\text{B},00} + r_{\text{B},01}}$$
$$\hat{\lambda}_{t^*,\text{B},1}(x) \to \frac{\lambda_{\text{B},10} r_{\text{B},10} + \lambda_{\text{B},11} r_{\text{B},11}}{r_{\text{B},10} + r_{\text{B},11}}$$

in probability. The $r_{l,j}$ are expected long run duration rates such that $\frac{d_{t^*,l,j}}{t^*} \to r_{t^*,l,j}$. It can be shown that

$$r_{t^*,\text{B},00} = \frac{1}{4}\cdot\frac{1}{4} \qquad r_{t^*,\text{B},01} = \frac{1}{4}\cdot\frac{3}{4}$$
$$r_{t^*,\text{B},10} = \frac{3}{4}\cdot\frac{1}{4} \qquad r_{t^*,\text{B},11} = \frac{3}{4}\cdot\frac{3}{4}$$

because they each depend on the probability of the homogeneous Poisson process A having events in the intervals $t-(0,1]$ and $t-(1,2]$ as $t \to \infty$. Substituting the given values of $\lambda_{l,j}$ and these values of $r_{t^*,l,j}$ into the m.l.e.s yields

$$\hat{\lambda}_{t^*,\text{B},0}(x) \to 3 \qquad \hat{\lambda}_{t^*,\text{B},1}(x) \to 3$$

That is, asymptotically, the dynamics of the process P cause the durations $d_{t^*,\text{B},00}, d_{t^*,\text{B},01}, d_{t^*,\text{B},01}$ and $d_{t^*,\text{B},11}$ to eventually balance out, so that the adding the edge $\text{A} \to \text{B}$ does not lead to an increase in BIC score. As a result, greedy search fails. □

In general, we say that P exhibits *detailed balance* if its dynamics are such that asymptotically, the durations of parent count vectors balance so that essential distinctions in P are obscured in a projected model. To characterize the essential distinctions in a m.p.p. P we define the set $\mathcal{O}^*(P)$ of *essential operators* of P as follows:

Definition 6. Let P be s.t. $\exists M \in \mathcal{M} : P \in \text{Markov}(M)$. Then the *essential operators* $\mathcal{O}^*(P)$ are the operators that take M_0 to the minimal model $M^*(P)$.

We now define detailed balance as follows:

Definition 7. Let $M \in \mathcal{M}$ be a model that makes an essential distinction in P. That is, let with $M = O(M')$ for some $O \in \mathcal{O}^*(P), M' \in \mathcal{M}$. P exhibits *detailed balance* if there exist $j_1, j_2 \in C_l$ for some $l \in \mathcal{L}$ so that $\pi(j_1) = \pi(j_2)$ and

$$\frac{\sum_{\bar{j}_1 \in \pi^{-1}(j_1)} \bar{\lambda}_{l,\bar{j}_1} r_{l,\bar{j}_1}}{\sum_{\bar{j}_1 \in \pi^{-1}(j_1)} r_{l,\bar{j}_1}} = \frac{\sum_{\bar{j}_2 \in \pi^{-1}(j_2)} \bar{\lambda}_{l,\bar{j}_2} r_{l,\bar{j}_2}}{\sum_{\bar{j}_2 \in \pi^{-1}(j_2)} r_{l,\bar{j}_2}}$$

where $\{\bar{\lambda}_{\bar{j}}\}$ is the parametrization of P in some $\bar{M} \geq M$.

This condition is analogous to faithfulness in Bayesian networks. We conjecture that, analogously to that case (Meek, 1995), the set of parameter values that exhibit detailed balance is of measure zero, as in the example below:

Algorithm 2 ForwardSearch()

$M \leftarrow M_0$
repeat
  refined $\leftarrow$ **false**
  $\mathcal{O} \leftarrow \{O : O(M) \succ M\}$
  for all $O \in \mathcal{O}$ **do**
    if $\mathcal{S}(O(M)) - \mathcal{S}(M) > 0$ **then**
      $M \leftarrow O(M)$ refined $\leftarrow$ **true**
    end if
  end for
until not refined
return M

Example 4. In Example 3, P exhibits detailed balance. If any one of λ_{A}, $\lambda_{\mathrm{B},00}$, $\lambda_{\mathrm{B},01}$, $\lambda_{\mathrm{B},10}$ or $\lambda_{\mathrm{B},11}$ is changed, P does not exhibit detailed balance.

For m.p.p.s that that are representable as RTGEMs and do not exhibit detailed balance , ForwardSearch() followed by BackwardSearch() is a finite consistent model selection procedure:

Theorem 9. *Suppose that $M \in \mathcal{M}$ and let P be such that $M = M^*(P)$. Let $\hat{M}_{t^*}$ be the result of* BackwardSearch(ForwardSearch()) *at time t^*. Then, if P does not exhibit detailed balance with respect to any model, $P(\hat{M}_{t^*} = M) \to 1$ as $t^* \to \infty$.*

It follows immediately from Theorem 2 that this procedure also gives consistent predictions of the future:

Theorem 10. *Suppose P is such that the conditions of Theorem 9 hold. Then, for any $\epsilon > 0, \Delta > 0$,*

$$P\left(\left|\log \frac{p(x_{t^*+\Delta}|x_{t^*})}{\hat{p}(x_{t^*+\Delta}|x_{t^*})}\right| > \epsilon\right) \to 0$$

as $t^* \to \infty$.

6 Conclusions

We have shown a universal approximating model family for bounded non-deterministic non-explosive finite-horizon smooth marked point processes. We have also presented a constructive proof of learnability for these models, by showing the asymptotic consistency of a greedy BIC structure learning procedure together with ML parameter estimates. In particular, our theoretical results show that the dependency structure of a universal family of point process models can be learned from data. We also show the predictive consistency of our learned models.

Our consistency results are quite general in a number of respects. We do not assume access to i.i.d. realizations of the process. We only assume that a single realization is observed for sufficiently long. While we assume the existence of bounds on the intensities and on the temporal range of dependencies, we do not assume prior knowledge of these bounds.

A number of interesting open questions remain. We conjecture that our predictive consistency result holds even when the true process is a bounded non-deterministic non-explosive finite-horizon smooth marked point processes that is not in our model class, but have not shown this in this paper. Another open question is whether structural consistency results analogous to ours exist in the the general case where the data is generated from a point process that is not an RTGEM. Indeed, it is unclear how best to formalize the concept of structural consistency in this case. While we only treat finite-horizon processes, other processes with limited history are also of interest. For example, a process with a binary latent variable can store a minimal amount of history for arbitrarily long. Results such as ours that apply to such processes would be of interest. Finally, stronger consistency results that give explicit learning rates would also be of interest.

References

R. P. Adams, I. Murray, and D. J. C. MacKay. Tractable nonparametric Bayesian inference in poisson processes with Gaussian process intensities. In *ICML*, pages 9–16, 2009.

P. Bühlmann and A. J. Wyner. Variable length Markov-chains. *Ann. Stat.*, 27(2):480–513, 1999.

D. M. Chickering and C. Meek. Finding optimal Bayesian networks. In *UAI*, 2002.

J. D. Cook. Error in normal approximation to the Poisson distribution. `http://www.johndcook.com/normal_approx_to_poisson.html`, 2013. Accessed Feb 25, 2013.

T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Thry.*, 13(1):21–27, 1967.

I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inf. Thry.*, 52(3):1007–1016, Mar. 2006.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Elementary Theory and Methods*, volume I. Springer, 2nd edition, 2003.

N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha. Scalable influence estimation in continuous-time diffusion networks. In *NIPS*, pages 3147–3155, 2013.

A. Golightly and D. J. Wilkinson. Bayesian sequential inference for stochastic kinetic biochemical network

models. *Journal of Computational Biology*, 13(3): 838–851, 2006.

A. Gunawardana, C. Meek, and P. Xu. A model for temporal dependencies in event streams. In *NIPS*, 2011.

K. Hornik. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

W. Lian, R. Henao, V. Rao, J. E. Lucas, and L. Carin. A multitask point process predictive model. In *ICML*, pages 2030–2038, 2015.

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *UAI*, 1995.

U. Nodelman, C. R. Shelton, and D. Koller. Continuous time Bayesian networks. In *UAI*, 2002.

A. P. Parikh, A. Gunawardana, and C. Meek. Conjoint modeling of temporal dependencies in event streams. In *UAI Workshop on Bayesian Modeling Applications*, 2012.

Z. Qin and C. R. Shelton. Auxiliary gibbs sampling for inference in piecewise-constant conditional intensity models. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 2015.

V. Rao and Y. W. Teh. Fast MCMC sampling for Markov jump processes and extensions. *JMLR*, 14: 3207–3232, 2013.

J. Rissanen. A universal data compression system. *IEEE Trans. Inf. Thry.*, IT-29(5):656–664, Sept. 1983.

J. Rissanen and G. G. Langdon, Jr. Universal modeling and coding. *IEEE Trans. Inf. Thry.*, IT-27(1): 12–23, Jan. 1981.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *JMLR*, 2001.

C. J. Stone. Consistent nonparametric regression. *Ann. Stat.*, 5(4):595–620, 1977.

M. J. Weinberger, J. J. Rissanen, and M. Feder. A universal finite memory source. *IEEE Trans. Inf. Thry.*, 41(3):643–652, May 1995.

J. C. Weiss and D. Page. Forest-based point process for event prediction from electronic health records. In *ECML*, 2013.

K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional Hawkes processes. *JMLR W&CP*, 28(3):1301–1309, 2013.

A Proofs

A.1 Universal Approximation

Theorem 1. *Suppose P is a stationary m.p.p. whose conditional intensity functions $\lambda_l(t|h)$ are bounded above and bounded away from zero, are Lipschitz continuous in t and h (i.e. in t and in the times of the events in h), and that there exits $t_h \geq 0$ such that $\lambda_l(t|h)$ does not depend on events in h at times earlier than $t - t_h$. Then, for any $\epsilon > 0$ there exists an RTGEM $\hat{M} \in \mathcal{M}$ and $\hat{P} \in Markov(\hat{M})$ such that $\frac{1}{t^*} D(P_{t^*} || \hat{P}_{t^*}) < \epsilon$.*

Proof. Let P be a m.p.p. that satisfies the conditions of the theorem. Thus, there exists t_h s.t. $\lambda_l(t|h) = \lambda_l(t|[h]_{[t-t_h,t)})$, where $[h]_{[t-t_h,t)}$ is set of events in h with times in $[t-t_h, t)$, there exist $\Lambda = \sup_{l,t,h} \lambda_l(t|h)$ and $\check{\lambda} = \inf_{l,t,h} \lambda_l(t|h) > 0$, and there exists K s.t. the conditional intensities ($\lambda_l(t|h)$ are K-Lipschitz continuous.

For fixed $\delta > 0$ and $m > 0$ we will define an RTGEM $\hat{M}$ and $\hat{P} \in \hat{M}$ that depend on δ and m, and then show that for any $\epsilon > 0$, we can choose δ so that $\frac{1}{t^*} D(P_{t^*} || \hat{P}_{t^*}) < \epsilon$. We choose $\hat{M}$ to be any RTGEM whose timescales for all labels have horizon $\hat{t_h}$ greater than t_h and all intervals smaller than δ. To specify $\hat{P} \in \hat{M}$ we need to specify $\hat{\lambda}_{l\hat{j}}$ for every label l and parent count vector $\hat{j}$. To do so, for each $l \in \mathcal{L}$, we define the set $A_{m,\delta}$ to be the set of all pairs (h,t) s.t.$|[h]_{[t-t_h-2\delta,t)}| < m$, and s.t. the time increments $t - t_i, t_{i-1} - t_{i-2}, \cdots, t_{i-m+1} - t_{i-m}$ all exceed δ. Then, choose $\hat{P} \in \text{Markov}(\hat{M})$ by setting $\hat{\lambda}$ according to $\hat{\lambda}_{l,\hat{j}} = \sup_{(h,t) \in A_{m,\delta} : \hat{c}_l(h,t) = \hat{j}} \lambda_l(t|h)$. We now have $\hat{\lambda}_l(t|h) \geq \lambda_l(t|h)$ on $A_{m,\delta}$.

We now show that for any $(h,t) \in A_{m,\delta}$, $\left|\lambda_l(t|h) - \hat{\lambda}_l(t|h)\right| < K\delta\sqrt{m}$. We represent h by $\{(t_1, l_1), \cdots, (t_i, l_i)\}$. By definition, $\hat{\lambda}_l(t|h) = \hat{\lambda}_{l,\hat{c}l(h,t)}$. Since this is defined via a supremum, for any $\epsilon > 0$, there exists a $(h', t') \in A_{m,\delta}$ with $\hat{c}_l(h', t') = \hat{c}_l(h,t)$ s.t. $0 \leq \hat{\lambda}_l(t|h) - \lambda_l(t'|h') < \epsilon$. Because $\lambda_l(\cdot|\cdot)$ depends only on the last m points in the sequence, the other points in h' can be assumed to match h. Since the count vectors $\hat{c}_l(h', t')$ and $\hat{c}_l(h,t)$ match, the inter-point distance exceeds δ and all timescale intervals are smaller than δ, the corresponding points in h and h' must have matching labels and fall in the same timescale intervals, and must therefore be at most δ apart. Thus, $||[t_1, \cdots, t_i, t]^T - [t'_1, \cdots, t'_i, t']^T|| < \delta\sqrt{m}$. Since we can find such (h', t') with $0 \leq \hat{\lambda}_l(t|h) - \lambda_l(t'|h') < \epsilon$ for all $\epsilon > 0$, we can find a sequence of such pairs $t^{(r)}, h^{(r)}$ s.t. $\lambda_l\left(t^{(r)}|h^{(r)}\right) \to \hat{\lambda}_l(t|h)$. Since $\lambda_l(t|h)$ is K-Lipschitz, $\left|\lambda_l(t|h) - \lambda_l\left(t^{(r)}|h^{(r)}\right)\right| < K\delta\sqrt{m}$. Taking the limit in r, $\left|\lambda_l(t|h) - \hat{\lambda}_l(t|h)\right| < K\delta\sqrt{m}$.

To bound the divergence, we expand $D(P_{t^*} || \hat{P}_{t^*})$ as

$$D(P_{t^*} || \hat{P}_{t^*}) = \mathbf{E}_{P_{|x_{t^*}|}}\left[\sum_{i=1}^{|x_{t^*}|} \mathbf{E}_{P_{x_{t^*}||x_{t^*}|}}\left[\sum_{l \in l} \mathbf{1}(l = l_i) \log \frac{\lambda_l(t_i|h_i)}{\hat{\lambda}_l(t_i|h_i)}\right]\right] + \sum_{l \in l} \mathbf{E}_{P_{|x_{t^*}|}}\left[\int_0^{t^*} \mathbf{E}_{P_{x_{t^*}||x_{t^*}|}}\left[\hat{\lambda}_l(\tau|h_\tau) - \lambda_l(\tau|h\tau)\right] d\tau\right]$$

where h_τ is the set of events in x_{t^*} with time less that τ.

Since the conditional intensities of P are bounded by Λ, the probability under P at any time τ that $[h]_{[\tau-t_h,\tau)}$ contains at least m events is bounded by this probability under a homogeneous Poisson process with intensity Λ, and therefore, by choosing m high enough, can be made arbitrarily close to zero uniformly in τ. For the same reason, given that there are less than m events in $[h]_{[\tau-t_h,\tau)}$, by choosing δ small enough, the probability that at least one inter-arrival time between them is less than than δ apart can also be made arbitrarily small, uniformly in τ. Thus, we can choose m and δ s.t. there is an arbitrarily small $\alpha_{m,\delta}$ so that $P((h,\tau) \notin A_{m,\delta}) < \alpha_{m,\delta}$.

When $(h,\tau) \in A_{m,\delta}$, we have that the first term in the divergence is negative and that the integrand in the second term is bounded by $K\delta\sqrt{m}$. Otherwise, the first term is bounded by $\mathbf{E}[|x_{t^*}|] \log \frac{\Lambda}{\check{\lambda}} \leq t^* \Lambda \log \frac{\Lambda}{\check{\lambda}}$, and the integrand in the second term is bounded by Λ. Thus, $D(P_{t^*} || \hat{P}_{t^*}) \leq \alpha_{m,\delta} t^* \Lambda \left(\log \frac{\Lambda}{\check{\Lambda}} + 1\right) + (1 - \alpha_{m,\delta}) t^* K\delta\sqrt{m}$. Given $\epsilon > 0$, we choose m and δ s.t. $\alpha_{m,\delta}\left(\Lambda \log \frac{\Lambda}{\check{\Lambda}} + \Lambda\right) < \frac{\epsilon}{2}$ and then further reduce δ until $K\delta\sqrt{m} < \frac{\epsilon}{2}$, so that $\frac{1}{t^*} D(P_{t^*} || \hat{P}_{t^*}) \leq \epsilon$. □

A.2 Parametric Consistency

Lemma 3. *For any P such that $\exists M \in \mathcal{M} : P \in Markov(\mathcal{M})$, and every $l \in \mathcal{L}, j \in C_l$, there exists $r_{l,j} > 0$ such that for every $\epsilon > 0, \delta > 0$ there exists $t > 0$ s.t. $P\left(\left|\frac{d_{t^*,l,j}(x)}{t^*} - r_{l,j}\right| < \delta\right) > 1 - \epsilon$ for $t^* > t$.*

Proof. At $t = 0$, all parent count vectors $c_l(h, 0)$ are equal to 0. Let $\tau_i(x)$ be the time of the ith return to this initial state. That is, let $\tau_i(x)$ be the ith time such that $c_l(h, \tau_i(x)) = 0$, $\lim_{t \uparrow \tau_i(x)} c_l(h,t)) \neq 0$.

By convention, $\tau_0 = 0$. Then, we define the ith episode as the sub-sequence $x_i(x) = \{(t - \tau_{i-1}, l) : (t, l) \in x, t \in [\tau_{i-1}(x), \tau_i(x))\}$ and the duration in episode i of parent count vector j of label l as $\Delta_{i,l,j}(x) = \sum_{k=1}^{|x_i|} \int_{t_{k-1}(x_i)}^{t_k(x_i)} \mathbf{1}(c_l(h_k(x_i), t') = j)dt'$. The episodes $\{x_i\}$ are then i.i.d., as are their total durations $\{\tau_i(x) - \tau_{i-1}(x)\}$ and their parent count vector durations $\{\Delta_{i,l,j}(x)\}$. In particular, $\tau_i(x) - \tau_{i-1}(x)$ are distributed as the r.v. $\tau(x)$, and since all conditional intensities in P are bounded and positive, $\tau(x)$ is upper and lower bounded by r.v.s with finite, positive, means and variances. $\tau(x)$ thus has finite, positive mean τ and finite variance. $\Delta_{i,l,j}(x)$ are distributed as the r.v. $\Delta_{l,j}(x)$, which is non-negative, and bounded above by τ. It therefore has finite mean $d_{l,j}$ and finite variance. Since the conditional intensities are bounded and positive, for any $l \in \mathcal{L}, j \in C_l$, there is positive probability that $\Delta_{1,l,j}(x) > 0$. Therefore, $d_{l,j} > 0$. We have that $d_{m,l,j}(x) = \sum_{i=1}^{m} \Delta_{i,l,j}(x)$. Thus, the central limit theorem holds for both $\tau_m(x)$ and $d_{m,l,j}(x)$, and therefore, for every $\epsilon' > 0$ and $\delta' > 0$ there exists m' such that for all $m > m'$, $P\left(\left(\left|\frac{d_{m,l,j}(x)}{m} - d_{l,j}\right| < \delta'\right) \wedge \left(\left|\frac{\tau_m(x)}{m} - \tau\right| < \delta'\right)\right) > 1 - \epsilon'$ Because $\tau > 0$, this implies that for every $\epsilon > 0$ and $\delta > 0$ there exists $\tilde{m}$ such that for all $m > \tilde{m}$,

$$P\left(\left|\frac{d_{m,l,j}(x)}{\tau_m(x)} - \frac{d_{l,j}}{\tau}\right| < \frac{\delta}{2}\right) > 1 - \frac{\epsilon}{4} \tag{2}$$

By Markov's inequality, for every $\epsilon > 0$, there exists $\bar{\tau}$ s.t. $P(\tau > \bar{\tau}) > 1 - \frac{\epsilon}{8}$. Since $\tau_{m+1}(x) - \tau_m(x)$ are i.i.d. distributed as τ, $P\left(\tau_{m+1}(x) - \tau_m(x) > \bar{\tau}\right) > 1 - \frac{\epsilon}{8}$. Thus, for a given m, the definition of $d_{m,l,j}(x)$ and $d_{t,l,j}(x)$ gives that for $t \in [\tau_m(x), \tau_{m+1}(x))$, $P\left(\frac{d_{t,l,j}(x)}{t} \in \left[\frac{d_{m,l,j}(x)}{\tau_m(x)+\bar{\tau}}, \frac{d_{m,l,j}(x)+\bar{\tau}}{\tau_m(x)}\right)\right) > 1 - \frac{\epsilon}{8}$. Since $\frac{d_{m,l,j}(x)}{\tau_m(x)} \in \left[\frac{d_{m,l,j}(x)}{\tau_m(x)+\bar{\tau}}, \frac{d_{m,l,j}(x)+\bar{\tau}}{\tau_m(x)}\right)$, this means

$$P\left(\left|\frac{d_{t,l,j}(x)}{t} - \frac{d_{t,l,j}(x)}{\tau_m(x)}\right| < \left|\frac{d_{m,l,j}(x)+\bar{\tau}}{\tau_m(x)} - \frac{d_{m,l,j}(x)}{\tau_m(x)+\bar{\tau}}\right|\right) > 1 - \frac{\epsilon}{8} \tag{3}$$

Notice that $\frac{d_{m,l,j}(x)+\bar{\tau}}{\tau_m(x)} - \frac{d_{m,l,j}(x)}{\tau_m(x)+\bar{\tau}} = \frac{d_{m,l,j}(x)}{\tau_m(x)}\left(\frac{d_{m,l,j}(x)+\bar{\tau}}{d_m(x)} - \frac{\tau_{m,l,j}(x)}{\tau_m(x)+\bar{\tau}}\right)$. From (2) and the from the central limit theorem on $\tau_m(x)$ and $d_{m,l,j}(x)$, for all $\delta > 0, \epsilon > 0$, there exists $\hat{m}$ s.t. for $m > \hat{m}$,

$$P\left(\left|\frac{d_{m,l,j}(x)+\bar{\tau}}{\tau_m(x)} - \frac{d_{m,l,j}(x)}{\tau_m(x)+\bar{\tau}}\right| < \frac{\delta}{2}\right) > 1 - \frac{\epsilon}{8} \tag{4}$$

Combining (3) and (4) we get that for all $\delta > 0, \epsilon > 0$, there exists $\hat{m}$ s.t. for $m > \hat{m}$,

$$P\left(\left|\frac{d_{t,l,j}(x)}{t} - \frac{d_{t,l,j}(x)}{\tau_m(x)}\right| < \frac{\delta}{2}\right) > 1 - \frac{\epsilon}{4} \tag{5}$$

Combining (2) and (5), we get that for all $\delta > 0, \epsilon > 0$, there exists $\hat{m}$ and $\tilde{m}$ s.t. for $t > \tau_{\max(\hat{m},\tilde{m})}(x), P\left(\left|\frac{d_{t,l,j}(x)}{t} - \frac{d_{l,j}}{\tau}\right| < \delta\right) > 1 - \frac{\epsilon}{2}$ From the central limit theorem on $\tau_m(x)$, given m^*, there exists t s.t. $P(t > \tau_{m^*}(x)) > 1 - \frac{\epsilon}{2}$. Thus, for every $\epsilon > 0, \delta > 0$ there exists $t > 0$ s.t. for all $t^* > t$,

$$P\left(\left|\frac{d_{t^*,l,j}(x)}{t^*} - r_{l,j}\right| < \delta\right) > 1 - \epsilon$$

where $r_{l,j} = \frac{d_{l,j}}{\tau}$. □

Lemma 4. *For any m.p.p. P such that $\exists M \in \mathcal{M}$: $P \in$ Markov(M) and any $l \in \mathcal{L}, j \in C_l$, there exists a constant $K > 0$ s.t. for all $\epsilon > 0$,*

$$P\left(\left|\frac{n_{t^*,l,j}(x)}{d_{t^*,l,j}(x)} - \lambda_{l,j}\right| > \epsilon \middle| d_{t^*,l,j}(x) = \hat{d}_{t^*,l,j}\right) < 2\left(1 - \Phi\left(\epsilon\sqrt{\frac{\hat{d}_{t^*,l,j}}{\lambda_{l,j}}}\right)\right) + \frac{K}{\lambda_{l,j}\hat{d}_{t^*,l,j}}$$

where $\Phi(\cdot)$ is the standard normal CDF.

Proof. By observation of the likelihood, given $d_{t^*,l,j}(x)$, each $n_{t^*,l,j}(x)$ is distributed as a independent Poisson r.v. with intensity $\lambda_{t^*,l,j}\hat{d}_{t^*,l,j}$. To show concentration, we note that due to the independent increments property, a Poisson r.v. with intensity $\lambda_{t^*,l,j}\hat{d}_{t^*,l,j}$ can be written as the sum of m i.i.d. Poisson r.v.s with intensity $\frac{\lambda_{t^*,l,j}\hat{d}_{t^*,l,j}}{m}$ for arbitrary m. Applying the Berry-Esséen Concentration Theorem and taking the limit $m \to \infty$ gives the desired result. Details can be found in Cook (2013). □

A.3 Structure of RTGEMs

Proposition 5. *Let P be such that $\exists M \in \mathcal{M}$ for which $P \in$ Markov(M). Then, there exists a unique minimal $M^*(P) \in \mathcal{M}$ such that $P \in$ Markov$(M^*(P))$ and $M' \geq M^*(P)$ for all M' for which $P \in$ Markov(M').*

Proof. Existence of a minimal model follows directly from the definition of RTGEMs. To prove uniqueness of the minimal model, suppose M_1 and M_2 are distinct models such that $P \in$ Markov(M_1) and $P \in$ Markov(M_2) and such that they are both minimal. Let $\mathcal{O}^* = \mathcal{O}_{M_1} \cap \mathcal{O}_{M_2}$, and define M_* to be the model reached by taking operators in $\mathcal{O}^*$ from M_0. We will

show that $P \in \text{Markov}(M^*)$ which will prove uniqueness by contradiction.

Let $\mathcal{O}_3 = \mathcal{O}_{M_1} \cup \mathcal{O}_{M_2}$, and define M_3 to be the model reached by taking operators in $\mathcal{O}_3$ from M_0. We then have $M_3 > M_1 > M^*$ and $M_3 > M_2 > M^*$. Thus, $P \in \text{Markov}(M_3)$. Let λ^1, λ^2, and λ^3 be the parameter that represents P in M_1, M_2, and M_3 respectively. We will show that $P \in \text{Markov}(M_1)$ implies a set of equality constraints between λ^1 and λ^3, and that $P \in \text{Markov}(M_2)$ implies a set of equality constraints on between λ^2 and λ^3. By transitivity of equality, we will then show that λ^1 and λ^2 satisfy a set of equality constraints that allow P to be represented in M^*.

Fix a label l and a parent count vector value $j^* \in C_l^*$ in model M^*. For every $j_1 \in C_{1,l}$ and $j_2 \in C_{2,l}$ s.t. $\pi_*^1(j_1) = j^*$ and $\pi_*^2(j_2) = j^*$ there exists $j_3 \in C_{3,l}$ s.t. $\pi_1^3(j_3) = j_1$ and $\pi_2^3(j_3) = j_2$. This is because the operators $\mathcal{O}_{M_1} \setminus \mathcal{O}^*$ and $\mathcal{O}_{M_2} \setminus \mathcal{O}^*$ refine different components of $c_l^*(h,t)$.

Because $P \in \text{Markov}(M_1)$, we have that $\lambda^3_{l,j_3} = \lambda^1_{l,\pi_1^3(j_3)}$. Similarly, because $P \in \text{Markov}(\hat{M})$, $P \in \text{Markov}(M_2)$, we have that $\lambda^3_{l,j_3} = \lambda^2_{l,\pi_2^3(j_2)}$. Thus, we have $\lambda^1_{l,\pi_1^3(j_3)} = \lambda^2_{l,\pi_2^3(j_2)}$ From above, there exists a j_3 to connect every j_1 and j_2 with $\pi_*^1(j_1) = \pi_*^2(j_2) = j^*$. Thus, by transitivity of equality, all the λ^1_{l,j_1} and λ^2_{l,j_2} for such j_1 and j_2 are equal, and we can set λ^*_{l,j_*} to this value, yielding a λ^* that shows $P \in \text{Markov}(M^*)$, which is the desired contradiction. □

A.4 Structural Consistency

Theorem 7. *Suppose $M \neq M' \in \mathcal{M}$ with $P \in$ Markov(M) and with either $P \notin$ Markov(M') or $M' > M$. Then,*

$$\lim_{t^* \to \infty} P\left(\mathcal{S}_{t^*}(M) > \mathcal{S}_{t^*}(M')\right) = 1$$

In particular, this holds for any $M' \neq M$ when $M = M^*(P)$.

Proof. Due to limited space we only present the proof for the cases $M > M'$ and $M' > M$. The general proof is slightly more cumbersome, but is analogous.

In the case $M > M', P \in \text{Markov}(M)$, and $P \notin \text{Markov}(M')$ Lemma 4 gives that $n_{t^*,l,j}(x)/d_{t^*,l,j}(x) \to \lambda_{l,j}$ and Lemma 3 gives that $d_{t^*,l,j}(x)/t^* \to r_{l,j}$. Using Proposition 6 we therefore have

$$\frac{\mathcal{S}_{t^*}(M) - \mathcal{S}_{t^*}(M')}{t^*} \to \sum_{l \in \mathcal{L}} \sum_{j' \in C_l'} \sum_{j \in \pi^{-1}(j')} r_{l,j} \lambda_{l,j} \log \lambda_{l,j} / \frac{\sum_{j \in \pi^{-1}(j')} \lambda_{l,j} r_{l,j}}{\sum_{j \in \pi^{-1}(j')} r_{l,j}}$$

in probability. By the log-sum inequality, the limit is strictly positive unless $\lambda_{l,j_1} = \lambda_{l,j_2}$ for every $l \in \mathcal{L}, j' \in C_l'$, and $j_1 \neq j_2 \in C_l$. If this were true, then $P \in \text{Markov}(M')$, which would be a contradiction. Therefore, $\lim_{t^* \to \infty} P\left(\mathcal{S}_{t^*}(M) > \mathcal{S}_{t^*}(M')\right) = 1$.

For the case $M' > M, P \in \text{Markov}(M)$, using a second order Taylor expansion of the logarithm yields

$$\mathcal{S}_{t^*}(M) - \mathcal{S}_{t^*}(M') \approx K \log t^* - \frac{1}{2} \sum_{l \in \mathcal{L}} \sum_{j \in C_l} \sum_{j' \in \pi^{-1}(j)} \frac{d_{t^*,l,j'}(x)}{n_{t^*,l,j'}(x)/d_{t^*,l,j'}(x)} \times \left(\frac{n_{t^*,l,j}(x)}{d_{t^*,l,j}(x)} - \frac{n_{t^*,l,j'}(x)}{d_{t^*,l,j'}(x)} \right)^2$$

for some $K > 0$. By Lemma 4, we have that for any $\epsilon > 0$, there exist a constant $A_\epsilon > 0$ and $d > 0$ such that for all $d_{t^*,l,j'}(x) > d, P\left(\left| \frac{n_{t^*,l,j}(x)}{d_{t^*,l,j}(x)} - \lambda_{l,j} \right| > \frac{1}{2} \sqrt{\frac{A_\epsilon}{d_{t^*,l,j}(x)}} \mid d_{t^*,l,j}(x) \right) < \frac{\epsilon}{4}$ and $P\left(\left| \frac{n_{t^*,l,j'}(x)}{d_{t^*,l,j'}(x)} - \lambda_{l,j'} \right| > \frac{1}{2} \sqrt{\frac{A_\epsilon}{d_{t^*,l,j'}(x)}} \mid d_{t^*,l,j'}(x) \right) < \frac{\epsilon}{4}$. By Lemma 3, there exists $t > 0$ s.t. for $t^* > t$, all $\hat{d}_{t^*,l,j'} > d$ with probability at least $1 - \frac{\epsilon}{4}$. Therefore, for $t^* > t$,

$$P\left(\left(\frac{n_{t^*,l,j}(x)}{d_{t^*,l,j}(x)} - \frac{n_{t^*,l,j'}(x)}{d_{t^*,l,j'}(x)} \right)^2 < \frac{A_\epsilon}{d_{t^*,l,j'}(x)} \right) > 1 - \epsilon$$

Thus, for all $\epsilon > 0$, there exists $t > 0$ s.t. for all $t^* > t$, $P\left(\mathcal{S}_{t^*}(M) - \mathcal{S}_{t^*}(M') > 0\right) > 1 - \epsilon$, or $P\left(\mathcal{S}_{t^*}(M) > \mathcal{S}_{t^*}(M')\right) \to 1$. □

Theorem 9 follows from Theorem 7 and the following:

Lemma 11. *Suppose P is representable in $\mathcal{M}$ and does not exhibit detailed balance with respect to any model. Suppose $M, M' \in \mathcal{M}$ are such that $M = O(M')$ for $O \in \mathcal{O}^*(P)$. Then*

$$\lim_{t^* \to \infty} P\left(\mathcal{S}_{t^*}(M) > \mathcal{S}_{t^*}(M')\right) = 1$$

Proof. Let $\bar{M} \in \mathcal{M}$ be the result of applying the operators $O^*(P) \setminus \mathcal{O}_M$ to M, so that $P \in \bar{M} \geq M$. Proposition 6 along with Lemmas 3 and 4 give

$$\frac{\mathcal{S}_{t^*}(M) - \mathcal{S}_{t^*}(M')}{t^*} \to \sum_{l \in \mathcal{L}} \sum_{j' \in C_l'} \sum_{j \in \pi^{-1}(j')} \left(\sum_{\bar{j} \in \pi^{-1}(j)} \lambda_{l,\bar{j}} r_{l,\bar{j}} \right) \times \log \frac{\sum_{\bar{j} \in \pi^{-1}(j)} \lambda_{l,\bar{j}} r_{l,\bar{j}}}{\sum_{\bar{j} \in \pi^{-1}(j)} r_{l,\bar{j}}} \Big/ \frac{\sum_{\bar{j} \in \pi^{-1}(j')} \lambda_{l,\bar{j}} r_{l,\bar{j}}}{\sum_{\bar{j} \in \pi^{-1}(j')} r_{l,\bar{j}}}$$

in probability. By the log-sum inequality, this is positive unless for every $l \in l'$ and $j_1 \neq j_2 \in C_l$ s.t. $\pi(j_1) =$

$\pi(j_2), \frac{\sum_{\bar{j}_1 \in \pi^{-1}(j_1)} \bar{\lambda}_{l,\bar{j}_1} r_{l,\bar{j}_1}}{\sum_{\bar{j}_1 \in \pi^{-1}(j_1)} r_{l,\bar{j}_1}} = \frac{\sum_{\bar{j}_2 \in \pi^{-1}(j_2)} \bar{\lambda}_{l,\bar{j}_2} r_{l,\bar{j}_2}}{\sum_{\bar{j}_2 \in \pi^{-1}(j_2)} r_{l,\bar{j}_2}}$,which contradicts the assumption that P does not exhibit detailed balance. □