Score-Based Learning of Graphical Event Models with Background Knowledge Augmentation

Debarun Bhattacharjya[1], Tian Gao[1], Dharmashankar Subramanian[1], Xiao Shou[2,3]

[1] Research AI, IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[2] Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA
[3] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA
{debarunb, tgao, dharmash}@us.ibm.com, shoux@rpi.edu

Abstract

Graphical event models (GEMs) are representations of temporal point process dynamics between different event types. Many real-world applications however involve limited event stream data, making it challenging to learn GEMs from data alone. In this paper, we introduce approaches that can work together in a score-based learning paradigm, to augment data with potentially different types of background knowledge. We propose novel scores for learning an important parametric class of GEMs; in particular, we propose a Bayesian score for leveraging prior information as well as a more practical simplification that involves fewer parameters, analogous to Bayesian networks. We also introduce a framework for incorporating easily assessed qualitative background knowledge from domain experts, in the form of statements such as 'event X depends on event Y' or 'event Y makes event X more likely'. The proposed framework has Bayesian interpretations and can be deployed by any score-based learner. Through an extensive empirical investigation, we demonstrate the practical benefits of background knowledge augmentation while learning GEMs for applications in the low-data regime.

1 Introduction & Related Work

Graphical event models (GEMs), also known as local independence graphs, are probabilistic graphical models for *temporal point processes* (TPPs) (Didelez 2008; Meek 2014). Unlike graphical representations for discrete-time dynamical models such as dynamic Bayesian networks and time series graphs (Dean and Kanazawa 1989; Murphy 2002; Eichler 1999), GEMs represent continuous-time dynamics where different types of events can occur irregularly over time.

Various types of GEMs have been proposed in prior work, differing in the parametric assumptions of how conditional intensity rates for each type of event vary as a function of historical occurrences. The most popular categories are those similar to Hawkes processes where intensities jump and decay over time (Zhou, Zha, and Song 2013b; Luo et al. 2015; Etesami et al. 2016; Xu, Luo, and Zha 2017), and those where intensities are piece-wise constant over time (Gunawardana, Meek, and Xu 2011; Bhattacharjya, Subramanian, and Gao 2018; Bhattacharjya, Gao, and Subramanian 2020). Among the latter family, *proximal graphical event models* (PGEMs) are simple yet effective at fitting many real world datasets (Bhattacharjya, Subramanian, and Gao 2018). Piece-wise constant GEMs have been shown to be a universal approximator for TPPs (Gunawardana and Meek 2016). There is also a burgeoning stream of recent work on neural architectures for learning TPP models when one has access to substantial event data (Du et al. 2016; Xiao et al. 2017; Mei and Eisner 2017; Omi, Ueda, and Aihara 2019; Shchur, Biloš, and Günnemann 2019; Gao et al. 2020; Zuo et al. 2020). This line of work has primarily focused on predicting the next type of event and its occurrence time.

We consider a separate but practically important research thread: how to learn GEMs in the *low-data regime*. This direction is currently understudied; a notable exception is some recent work on a new parametric family for TPPs based on expert-provided temporal logical rules (Li et al. 2020). In contrast, we propose a framework for learning a specified parametric GEM (such as PGEM) through score-based approaches while augmenting data with knowledge. Furthermore, we enable incorporation of various forms of knowledge, including simple qualitative statements that are easy to assess or readily available, such as 'event X depends (or does not depend) on event Y' or 'event Y makes event X more (or less) likely'. Note that the former statement cannot be represented by first order temporal logic as in Li et al. (2020). Other prior work considers knowledge as a data-dependent measure and hence may be unsuitable for low-data problems (Zhang, Sharma, and Liu 2021).

In this work, we are particularly interested in recovering the underlying model from limited data, i.e. parameter estimation and structure discovery. The latter task is notoriously challenging in dynamical models due to time dependent confounding (Keiding 1999), and learning an underlying GEM structure can require much more event data than i.i.d data requirements for Bayesian networks. It is therefore important in practical applications to enable current GEM learning approaches to effectively exploit any background knowledge as an inductive bias whenever it is available. However, a practical learner should ideally be able to override any initial knowledge given enough data; this is particularly important when the knowledge is inconsistent with the ground truth.

There is a long line of prior research on incorporating qualitative knowledge in graphical models such as Bayesian networks (Lucas 2005; Feelders 2007), mostly around parameter estimation given the graph (Wittig and Jameson 2000; Liao and Ji 2009; Zhou, Fenton, and Neil 2014). We extend some of these conceptual ideas from learning joint probability distributions towards the more complex endeavor of learning TPPs as represented by families of GEMs. Specifically, we present a framework where background knowledge from qualitative statements in the form of soft quantitative constraints about temporal dynamics can augment data to guide a score-based learning approach.

Contributions. We make the following contributions:

- We propose a novel Bayesian score for the piecewise-constant GEMs family, showing a simplification that could be practical as it requires only 1 or 2 assessments.
- We describe a general *incompatibility framework* for incorporating background knowledge while learning GEMs by deploying any score-based learner.
- In one instance of the framework, we show how to incorporate directional statements resulting in inequalities for improving parameter estimation in PGEMs. In another instance, we show how to incorporate underlying process statements for enhancing structure discovery in GEMs.
- We conduct an extensive empirical investigation to demonstrate the practical benefits of background knowledge augmentation on various tasks while learning PGEMs for low-data regime applications.

2 Background

Event Datasets. An event dataset has time-stamped streams of events of different types. Formally, it is denoted $\mathcal{D} = \{\mathcal{D}_k\}_{i=1}^{K}$, where $\mathcal{D}_k = \{(l_i^k, t_i^k)\}_{i=1}^{N_k}$ and event label (or type) l_i^k belongs to a known label set (or alphabet), $l_i^k \in \mathcal{L}$, such that the number of event labels $|\mathcal{L}| = M$. Time stamp t_i^k is the occurrence time of the i^{th} event in the k^{th} stream, $t_i^k \in \mathbb{R}^+$, assumed temporally ordered between start time $t_0^k = 0$ and final time $t_{N+1}^k = T^k$ which could differ across event streams. There are K streams of events in the dataset with a total event count of $N = \sum_{k=1}^{K} N_k$ events, and a total time horizon of $T^* = \sum_{k=1}^{K} T^k$.

Temporal Point Processes (TPPs). Event datasets can be viewed as samples from a multivariate TPP associating each label in $\mathcal{L}$ with a counting process (Aalen, Borgan, and Gjessing 2008; Daley and Vere-Jones 2002). Prior work uses a Doob-Meyer decomposition to show that a conditional intensity function for measuring the rate at which an event label occurs is sufficient to characterize TPPs under general assumptions. The conditional intensity for an event label X at any time t is a function of historical event occurrences, i.e. $\lambda_x(t|h_t)$ where h_t includes all events up to time $t, h_t = \{(l_i, t_i) : t_i < t\}$. Due to intractability in representation and learning, TPP models need to assume specific forms for historical dependence of conditional intensity rates.

Graphical Event Models (GEMs). GEMs (Didelez 2008; Meek 2014; Gunawardana and Meek 2016) are a graphical representation for TPPs that capture *process independence* relationships among different types of events, which is a notion of independence pertaining to temporal dynamics (Schweder 1970; Didelez 2008), analogous to conditional independence in Bayes nets (Pearl 1988). Informally, for event labels $X, Y, \mathbf{Z} \subset \mathcal{L}$ s.t. $Y \cap \mathbf{Z} = \emptyset$, X is process independent of Y given $\mathbf{Z}$, denoted $Y \not\rightarrow X|\mathbf{Z}$, when label X has a conditional intensity such that if the historical occurrences of label set $\mathbf{Z}$ are known, then those of label Y do not provide any further information. Every event label X is process independent of its non-parents given its parents $\mathbf{U}_X$ in the GEM graph $\mathcal{G}$. This implies that the dependence of X's conditional intensity on historical occurrences simplifies as $\lambda_x(t|h_t) = \lambda_x(t|[h(\mathbf{U}_X)]_t)$, where $\mathbf{U}_X$ are X's parents and $[h(\mathbf{U}_X)]_t$ is the history restricted to labels in set $\mathbf{U}_X$. One can identify process (in)dependence relations from a directed and possibly cyclic GEM graph using a separation criterion analogous to d-separation in Bayes nets (Didelez 2008; Mogensen, Malinsky, and Hansen 2018).

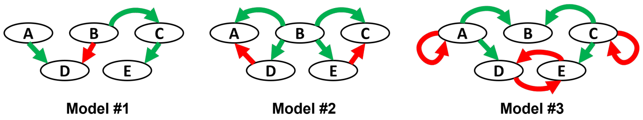

Figure 1: Graphs of 3 example PGEMs, used later for experiments with synthetic data. Green (red) arcs indicate amplification (inhibition) effects, i.e. when a parent increases (decreases) a child's conditional intensity rate.

Proximal Graphical Event Models (PGEMs). PGEMs belong to the afore-mentioned broad family of graphical models that characterize relationships between various types of events (Bhattacharjya, Subramanian, and Gao 2018). Each node in a PGEM graph represents a point process for an event type where the rate of its occurrence depends only on whether or not its parents have occurred at least once in the most recent history, which is represented in the form of a set of learnable proximal windows corresponding to each edge in the graph. Fig. 1 depicts 3 PGEM graphs, each with 5 nodes. As an illustrative example, the rate at which D occurs at any time in model #1 depends on whether its parents A and B occur at least once in their respective proximal time windows. Thus, the process for D has 4 intensity rates: $\Lambda_d = \{\lambda_{d|a,b}, \lambda_{d|\bar{a},b}, \lambda_{d|a,\bar{b}}, \lambda_{d|\bar{a},\bar{b}}\}$, where $\lambda_{d|a,\bar{b}}$ denotes the rate when A occurs at least once but B does not occur in its proximal window. While learning a PGEM could be entirely data-driven, we make the case in this paper that knowledge augmentation could be particularly beneficial in the situation where data is limited.

Score-based Learning. Learners for parametric GEMs (mostly but not exclusively in the piece-wise constant family) are typically score-based approaches using a forward-backward graph search (Chickering 2002). The Bayesian information criterion (BIC) is a popular score across different models (Bhattacharjya, Subramanian, and Gao 2018; Bhattacharjya, Gao, and Subramanian 2020; Yu et al. 2020). Although a Bayesian approach to parameter estimation has been proposed previously for a piece-wise constant intensity model (PCIM) learner (Gunawardana, Meek, and Xu 2011), the underlying assumptions are not formally stated and the high number of required assessments are impractical. Here we propose a new Bayesian score for more general GEMs, including an important practical simplification.

3 The Bayesian Gamma Score

We consider a class of GEMs that we refer to as *piecewise constant graphical event models* (PCGEMs) and that generalize prior models. In this class, the model $\mathcal{M}$ consists of graph $\mathcal{G}$ and a known number of historical windows $\mathbf{W} = \{\mathbf{W}_X\}$ as well as conditional intensity parameters $\Lambda = \{\Lambda_X\}$ for every node X in the graph, whose parents in $\mathcal{G}$ are denoted $\mathbf{U}_X$. Given the parents and windows, there is a known mapping from any event history to a finite domain Σ_X for every node X, such that $\Lambda_X = \{\lambda_{x|\mathbf{s}}\}$ for all historical summaries $\mathbf{s} \in \Sigma_X$. (We hide the dependence of Σ_X on $\mathbf{U}_X$, $\mathbf{W_X}$ to avoid clutter.) PGEM is a special case of PCGEM with a single window and where mapping results in binary instantiations of parent combinations, thus every summary $\mathbf{s}$ is a parental instantiation $\mathbf{u}$ and $|\Sigma_X| = 2^{|\mathbf{U}_X|}$.

We wish to score a model $\mathcal{M}$ on how well it fits a dataset $\mathcal{D}$. The log of the joint distribution factorizes as:

$$\log P(\mathcal{D}, \mathcal{M}) = \log P(\mathcal{G}) + \log P(\mathbf{W}|\mathcal{G}) + \log P(\Lambda|\mathbf{W}, \mathcal{G}) + LL, \tag{1}$$

where LL is the log likelihood of the data, i.e. log of $P(\mathcal{D}|\Lambda, \mathbf{W}, \mathcal{G})$, which for a PCGEM is:

$$LL = \textstyle\sum_{X \in \mathcal{L}} LL_X = \sum_{X, \mathbf{s} \in \Sigma_X} \left(-\lambda_{x|\mathbf{s}} D(\mathbf{s}) + N(x, \mathbf{u}) \ln(\lambda_{x|\mathbf{s}})\right), \tag{2}$$

where $N(x, \mathbf{s})$ is the number of times that X is observed and the history maps to state $\mathbf{s}$ of Σ_X, $D(\mathbf{s})$ is the duration over which the history maps to state $\mathbf{s}$, and $\Lambda_X = \{\lambda_{x|\mathbf{s}}\}$ are intensities for X.

Assuming non-informative priors on windows given any graph $\mathcal{G}$, independence of conditional intensity priors as well as parent set priors, allows for two simplifications to Eq. (1): (1) the score becomes *decomposable*, i.e. the total score is a sum over all label-specific scores, and (2) it allows one to disregard windows, since maximizing Eq. (1) is equivalent to maximizing the sum of:

$$S_X = \log P(\mathbf{U}_X) + \log P(\Lambda_X|\mathbf{W}, \mathcal{G}) + LL_X, \tag{3}$$

over all nodes X. The following result generalizes the Bayesian updating approach from (Gunawardana, Meek, and Xu 2011) after recognizing that the Gamma distribution is a conjugate prior for intensities:

Theorem 1 *If conditional intensity priors are Gamma distributions and event dataset $\mathcal{D}$ is a complete TPP sample from the underlying PCGEM, then maximizing Eq. (3) is equivalent to maximizing the Bayesian Gamma (BG) score:*

$$BG_X = \log P(\mathbf{U}_X) + LL^*_X, \tag{4}$$

where LL^*_X is the log likelihood computed at the following conditional intensity estimates:

$$\hat{\lambda}_{x|\mathbf{s}} = \frac{N(x, \mathbf{s}) + N'(x, \mathbf{s}) - 1}{D(\mathbf{s}) + D'(\mathbf{s})}, \forall X, \mathbf{s}, \tag{5}$$

which includes counts of observing label X in summary state $\mathbf{s}$ in the dataset, $N(x, \mathbf{s})$, and durations over which that state is true, $D(\mathbf{s})$. $N'(x, \mathbf{s})$ and $D'(\mathbf{s})$ are effective counts and durations of the prior over conditional intensities.

The conditional intensity distributions are updated from the prior $\lambda^0_{x|\mathbf{s}} \sim$ Gamma $(N'(x, \mathbf{s}), D'(\mathbf{s}))$ to the posterior $\lambda_{x|\mathbf{s}} \sim$ Gamma $(N'(x, \mathbf{s}) + N(x, \mathbf{s}), D'(\mathbf{s}) + D(\mathbf{s}))$.

Theorem 2 *The BG score is score-equivalent, i.e. the scores of two graphs with the same process independence statements are identical.*

We highlight that the general BG score is onerous to assess, similar to the BD score (Heckerman, Geiger, and Chickering 1994) in Bayesian networks since it requires assessments for every label and summary state. One can reduce the burden significantly by making a strong simplifying assumption about the prior parameters. We refer to this version as the BGPi score due to additional assumptions (Poisson, identical) as specified next.

Theorem 3 *If the prior TPP dynamics are modeled as a multivariate Poisson process with identical intensity rates for all labels, then the conditional intensity estimates in Eq. (5) specialize as:*

$$\hat{\lambda}_{x|\mathbf{s}} = \frac{N(x, \mathbf{s}) + (N'/|\Sigma_X|) - 1}{D(\mathbf{s}) + (D'/|\Sigma_X|)}, \forall X, \mathbf{s}, \tag{6}$$

where D' is the a-priori effective duration of observing an effective N' number of events of any particular type.

For our experimental investigation using BGPi, we make a final simplification and set $N' = 1$, in which case one need only assess a single number D' representing the user's prior about the time to observe 1 event of a particular type. Besides this parameter, we model the parent set prior using a complexity term $\kappa \in (0, 1]$ assuming $P(\mathbf{U}_X) \propto \kappa^{|\mathbf{U}_X|}$. A lower κ controls complexity by penalizing larger parent sets; $\kappa = 1$ implies a uniform prior over any parent set.

In summary, the BGPi score is a special case of the BG score that can be deployed much more easily since it only involves two parameters: effective single event duration D' and optionally, complexity parameter κ. Furthermore, it can be used for any model in the broad PCGEM family.

4 Qualitative Knowledge

The BG score requires background knowledge in a form that some may find difficult to assess, due to its quantitative nature, even when simplified to a few parameters. In this section, we consider *qualitative knowledge* that might be easier to assess or more readily available. For instance, it might be possible in some applications to acquire qualitative knowledge such as pairwise influences between types of events from sources such as causal knowledge graphs (Sap et al. 2019; Heindorf et al. 2020; Hassanzadeh 2021).

We first provide a general framework for incorporating these forms of knowledge while learning event models, and then describe how the framework can be deployed for specific types of qualitative statements while learning PGEMs. For broader practical applicability, we assume that the knowledge might be incompatible with the ground truth, therefore we treat the knowledge as potentially fallible.

Incompatibility Framework

Suppose an expert (or group of experts) or some other source of knowledge provides a set of (potentially qualitative) statements S that can be mapped to a set of soft quantitative constraints C pertaining to a GEM $\mathcal{M}$. A function $f(C, \mathcal{M})$ that signifies the extent to which constraints C are *incompatible* with respect to the model $\mathcal{M}$ can be used to modify the log likelihood on data $\mathcal{D}$:

$$LL' = LL - w * f(C, \mathcal{M}), \tag{7}$$

where $w \in \mathbb{R}$ captures the strengths of their belief about the validity of constraints C.

If we choose $f(C, \mathcal{M}) = 0$ whenever C is compatible with $\mathcal{M}$ else otherwise positive, then the modified LL' above retains a log likelihood semantic. This is because the augmented knowledge can be seen as an additional observation whose likelihood is given by the probability that the expert says that the statements corresponding to constraints C are compatible with model $\mathcal{M}$ (Wittig and Jameson 2000). This probability is exp$(-w * f(C, \mathcal{M}))$; it is 1 when there is compatibility, and the weight w determines how quickly it goes to 0 with increasing violation $f(\cdot)$. This results in a Bayesian interpretation of Eq. (7) where a system with a uniform prior over models has a posterior that is proportional to LL', after observing data and the knowledge. We highlight another Bayesian interpretation of Eq. (7) in Appendix A[1] where the background knowledge can be considered incorporated in the prior, but the choice of interpretation does not affect computations in any way.

The weight in Eq. (7) can be treated as a model hyperparameter and therefore estimated from a validation set. Since a separate dataset may not be available in the low-data regime, the Bayesian interpretations could potentially inform a user about how to assess these weights. We discuss some guidelines briefly in Appendix B.

We refer to any objective function where $f(\cdot)$ is in $\mathbb{R}^+$ as *penalized log likelihood* since the log likelihood LL can only be decreased (and not increased) with the additional constraints. Note that any score-based approach can be used for learning by replacing LL with LL' from Eq. (7), for instance, the BIC score for a node X is:

$$BIC_X = LL_X - w * f(C_X, \mathcal{M}_X) - |\Lambda_X| \log T^*, \tag{8}$$

where LL_X is the log likelihood for X, C_X are constraints involving X, $\mathcal{M}_X$ includes all model components for X, $|\Lambda_X|$ is the # of parameters of X and T^* is the total time horizon. We propose specific types of constraints and incompatibility functions $f(\cdot)$ in subsequent sub-sections.

Parameter Inequalities

The proposed framework can be applied to inequality statements on conditional intensities for the same node X given its parents $\mathbf{U}_X$ and any other parameters involved in the GEM (such as windows for PGEMs). Consider a set of inequality statements I_X (indexed by i), each over a set of intensities denoted Λ_x^i: $g_i(\Lambda_x^i) = \sum_{\lambda_{x|\mathbf{u}} \in \Lambda_x^i} a_{\mathbf{u}}^i \lambda_{x|\mathbf{u}} + c^i \geq 0$, where coefficients $a_{\mathbf{u}}^i, c^i$ are real-valued constants. The following function can be used to measure incompatibility with a set of such inequality constraints: $f(I_X, \Lambda_X) = \left(\sum_{i \in I_X} -g_i(\Lambda_x^i))\right)^+$, where $f(\cdot)^+$ denotes $\max(0, f(\cdot))$. A net larger sum of differences in the intensities as provided by the inequalities results in more incompatibility.

While the above function allows for any such inequalities over X, we are particularly interested in qualitative statements of the form 'Y makes X more (or less) likely', which can easily be mapped to inequalities for a PGEM. We interpret such a statement to mean that when X has Y as a parent and when optionally there are other parents $\mathbf{Z}$, $\lambda_{x|y,\mathbf{z}} \geq$ (or $\leq$) $\lambda_{x|\bar{y},\mathbf{z}}$, $\forall$ binary instantiations $\mathbf{z}$ of $\mathbf{Z}$. Recall that y ($\bar{y}$) in the subscript refers to when the label Y does (does not) occur in the recent history.

A learner for maximizing the score, like BIC in Eq. (8), requires a subroutine to compute conditional intensities that maximize the penalized log likelihood LL'_X using the incompatibility function $f(I_X, \Lambda_X)$. In general, one can use any gradient-based approach to find the local optima. Note that it is possible to write a closed form expression for the gradient with respect to any intensity parameter.

The maximum likelihood estimates of the penalized log likelihood LL'_X have an intuitive interpretation, which we highlight using the following important case of a single inequality regarding two intensities. In the following, we denote $\hat{\lambda}_{x|\mathbf{u}} = \frac{N(x,\mathbf{u})}{D(\mathbf{u})}$ as the un-penalized maximum log likelihood estimates, where as usual $N(\cdot)$ and $D(\cdot)$ are the relevant summary statistics from data.

Theorem 4 *If a statement implies $\lambda_{x|\mathbf{u}} \geq \lambda_{x|\mathbf{u}'}$ and this pair of intensities is not involved in any other constraint, then the optimal intensities are as follows (subject to further non-negativity constraints):*

$$\lambda^*_{x|\mathbf{u}}, \lambda^*_{x|\mathbf{u}'} = \begin{cases} \hat{\lambda}_{x|\mathbf{u}}, \hat{\lambda}_{x|\mathbf{u}'}, & \textit{if } \hat{\lambda}_{x|\mathbf{u}} \geq \hat{\lambda}_{x|\mathbf{u}'} \\ \frac{N(x,\mathbf{u})}{D(\mathbf{u})-w}, \frac{N(x,\mathbf{u})}{D(\mathbf{u})+w}, & \textit{else if } w < w^* \\ \frac{N(x,\mathbf{u})}{D(\mathbf{u})-w^*}, \frac{N(x,\mathbf{u})}{D(\mathbf{u})+w^*}, & \textit{otherwise} \end{cases} \tag{9}$$

where w is the provided weight and w^* is the weight at which intensity estimates in the second case are identical.

The above result shows that the knowledge through the inequality has no effect when the estimated intensities from the data are consistent with the knowledge, otherwise the weight acts as an effective duration change that brings the pair of intensities closer together in the desired direction. In the general case, the optimal estimate for a region of the intensity space, as defined by the inequalities, modifies the un-penalized (knowledge free) maximum log likelihood estimates by effectively modifying the duration through an increment or decrement that multiplies the weight by the number of inequalities violated (when the weight w applies uniformly over all inequalities in I_X).

In this work, we assume there are no inequalities involving conditional intensities across different event labels, leaving more general cases such as this one or the case of statement-specific confidence (or weights) to future work.

[1] Appendices can be found in the arXiv version of the paper.

Process (In)dependence Statements

The proposed framework can also be applied to soft constraints involving the graphical structure of a GEM. Consider a set of process dependence or independence related constraints P_X for a node X (indexed by i), i.e. either of the form $Y \rightarrow X|\mathbf{Z}$ or $Y \not\rightarrow X|\mathbf{Z}$. Consider binary function $h_i(P_X^i, \mathbf{U}_X)$ which is 1 or -1 depending on whether the constraint is consistent or inconsistent with X's parents $\mathbf{U}_X$. The following function can then be used to measure the extent of incompatibility: $f(P_X, \mathbf{U}_X) = \left(\sum_i -h_i(P_X^i, \mathbf{U}_X)\right)^+$.

An important special case of this type of constraint is a qualitative statement such as 'X depends on Y', which can be mapped to the process dependence constraint $Y \rightarrow X|\emptyset$; this is equivalent to stating that Y is a parent of X in a GEM. We note that these constraints depend only on the graph and not on the parameters, and therefore apply to any GEM.

It is straightforward to incorporate the incompatibility function $f(P_X, \mathbf{U}_X)$ using any graph search approach (such as forward-backward search) for optimizing any score (such as BIC) as shown in Eq. (8). As mentioned previously, prior work has shown how to gauge process (in)dependence in GEMs using a graphical separation criterion.

5 Experiments

We conduct an extensive empirical investigation to study the advantages of our proposed methods, referring the reader to Appendices D and E for details about datasets and experimental settings respectively. We consider 3 synthetic PGEM datasets and 4 popular real-world event datasets: Diabetes, LinkedIn, Mimic and Stack Overflow.

Score Comparison. We compare various scores for the task of structure discovery using synthetic datasets generated from the 3 PGEMs with graphs as shown in Fig. 1. We generate $K = \{5, 10\}$ streams, each up to horizon $T = 1000$ and involving $\sim$ 1K to 1.5K events, and use a forward-backward search to compare the PGEM learner using the typical BIC score with the Akaike information criterion (AIC) as well as the Bayesian score BGPi. PGEM with the AIC score has not been empirically studied previously. We also include the following models:

- An implementation of CPCIM, a the piece-wise constant TPP (Parikh, Gunawardana, and Meek 2012).
- Additional baselines from the multi-dimensional Hawkes process literature (Eichler, Dahlhaus, and Dueck 2017), namely Hawkes with an exponential kernel (Hawkes-Exp) and ADM4 (Zhou, Zha, and Song 2013a) which learns an infectivity matrix where each entry i, j denotes the influence from event j to i. We threshold to obtain a binary matrix to compare with ground truth (Zhang and Yan 2021).
- CAUSE: A recent Granger causal inspired approach to recover the structure of a GEM (Zhang et al. 2020); we use the default settings in publicly available code. [2]

Table 1 compares F1 scores. Results for BGPi are shown at the optimal setting for hyper-parameters for each model, over the grid $D' = \{1, 3, 5, 10\}$ and $\kappa = \{0.01, 0.02, 0.03, 0.05, 0.1\}$. Let us first compare the three PGEM score-based learners with different scores. We see that BIC has high F1 scores for sparse graphs (like model #1) since it has excellent precision but fails to identify some parents for dense graphs. In contrast, AIC chooses more parents and has better F1 for denser graphs even though precision is often lower. However, it exhibits some undesirable behavior such as occasionally poorer performance with more data. Although BGPi's F1 scores are favorable here, we stress that Bayesian scores depend heavily on the choice of priors for small datasets, therefore hyper-parameter choices can be important. This is typical with all Bayesian scores, such as those for learning Bayesian networks (Silander, Kontkanen, and Myllymaki 2007). For example, a smaller κ such as 0.01 leads to better performance for sparser graphs since it adds inductive bias towards the ground truth structure. An advantage of scores such as BIC and AIC is that they do not require additional hyper-parameters. Further details around BGPi's sensitivity to D' and κ are provided in Appendix E.

Table 1 also reveals that the performance of non-PGEM learners is generally worse here. CPCIM learns graphs that are too sparse. The Hawkes models are clearly at a disadvantage since the synthetic data is generated from a PGEM. F1 scores for the neural-based CAUSE are also much lower than the top performing PGEM learners – but they do not vary at all with the chosen dataset sizes, indicating that these may be smaller datasets than such methods expect.

In subsequent experiments pertaining to the effects of incorporating qualitative knowledge, we use the BIC score exclusively for score-based learning of PGEMs, for simplicity of exposition. We note that the trends described also generally hold for other scores, as illustrated by recreating some of the figures using AIC and BGPi scores in Appendix E.

Directional Statements (Inequalities): Correct vs. Incorrect Statements. We consider the task of parameter estimation given a known structure, and study the effect of correct vs. incorrect directional statements about pairs of event labels. $K = \{1, 2, 5, 10\}$ event streams of horizon $T = 100$ were generated using model #3 from Fig. 1, and one directional statement (such as A makes A more likely to happen) was provided per node as qualitative knowledge.

Fig. 2 shows mean square error (MSE) between the estimated and true conditional intensities, averaged over the 5 labels, for both the case of correct and incorrect statements as gauged by whether the knowledge is consistent with the ground truth. A confidence level of $w = 0$ is equivalent to having no additional knowledge. We observe that although the statements could be helpful to reduce MSE for the very low data case ($K = 1$), they do not carry much information for parameter estimation as our choice of objective function prevents drastic changes in conditional intensity estimates.

Parental Statements (Process Dependence): Correct vs. Incorrect Statements. We revisit the synthetic data streams from the previous experiment with the same horizon and choices of # of streams K. Here we consider the structure discovery task using one parental statement (such as label A depends on A) per node as qualitative knowledge.

Fig. 3 shows F1 scores for the task of structure learn-

[2] https://github.com/razhangwei/CAUSE

	Model # 1		Model # 2		Model # 3	
	$K = 5$	$K = 10$	$K = 5$	$K = 10$	$K = 5$	$K = 10$
CAUSE	0.28 ± 0	0.28 ± 0	0.39 ± 0	0.39 ± 0	0.48 ± 0	0.48 ± 0
CPCIM	0.30 ± 0.40	0.16 ± 0.40	0.65 ± 0.10	0.72 ± 0.21	0.60 ± 0.13	0.74 ± 0.10
Hawkes-ADM4	0.69 ± 0.16	0.75 ± 0.10	0.58 ± 0.16	0.59 ± 0.12	0.55 ± 0.08	0.51 ± 0.05
Hawkes-Exp	0.61 ± 0.13	0.63 ± 0.13	0.58 ± 0.16	0.59 ± 0.13	0.51 ± 0.09	0.49 ± 0.09
PGEM-AIC	0.66 ± 0.06	0.65 ± 0.08	0.67 ± 0.17	0.73 ± 0.09	$\mathbf{0.93 \pm 0.07}$	$\mathbf{0.90 \pm 0.05}$
PGEM-BGPi	$\mathbf{0.97 \pm 0.06}$	$\mathbf{0.98 \pm 0.06}$	$\mathbf{0.70 \pm 0.12}$	$\mathbf{0.78 \pm 0.10}$	0.85 ± 0.08	$\mathbf{0.90 \pm 0.04}$
PGEM-BIC	0.89 ± 0.07	$\mathbf{0.98 \pm 0.01}$	0.58 ± 0.08	0.66 ± 0.15	0.66 ± 0.12	0.75 ± 0.06

Table 1: Mean F1 scores with error bars over 10 samples from learning the 3 PGEM graphs in Fig. 1 using $K = \{5, 10\}$ generated streams. We compare a PGEM learner with 3 scores (BIC, AIC and BGPi) as well as CAUSE, CPCIM and two versions of the multivariate Hawkes process (Hawkes-Exp and Hawkes-ADM4). Best mean results are shown in bold.

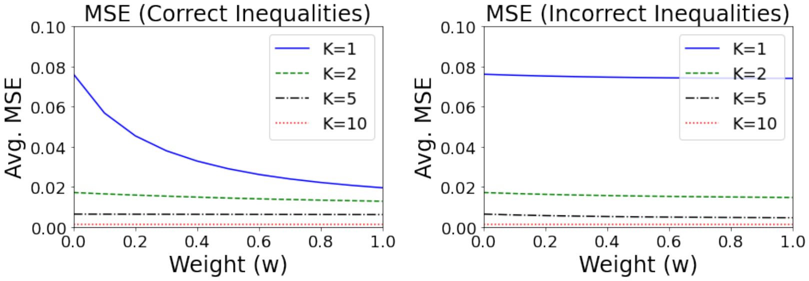

Figure 2: Avg. mean square error (MSE) over event labels as a function of confidence (w) about directional statements for model #3 in Fig 1, shown for number of data streams $K \in \{1, 2, 5, 10\}$. Left: Correct statements. Right: Incorrect statements.

ing, for both the case of correct as well as incorrect statements. From the left plot, we see that correct parental statements can greatly improve performance on structure discovery. For instance, for $K = 2$ streams, a high confidence in the statements can improve the F1 score from around 0.2 to almost 0.8. The potential improvement plateaus after a certain confidence level is attained since the knowledge is limited. As anticipated and desirable, the influence of the knowledge is reduced when more data is available. On the flip side, highly confident incorrect statements can hamper performance. However, since the knowledge isn't a hard constraint, more data allows one to regain performance.

Parental Statements (Process Dependence): Effect of Number of Statements. In this experiment, we study how confident correct parental statements can affect structure discovery performance when there is access to only limited data. We leverage the following 3 real-world datasets:

- **LinkedIn** (Xu, Luo, and Zha 2017): Employment related events such as joining a new role for 1000 LinkedIn users.
- **Mimic-II** (Saeed et al. 2011): Events corresponding to electronic health records from Intensive Care Unit patient visits over 7 years.
- **Stack Overflow** (Grant and Betts 2013): Events for engagement of 1000 users (chosen from Du et al. (2016)) around receipt of badges in a question answering website.

Although we do not know the true underlying structures for these datasets, we can estimate them since we have access to substantial data. We assume that a PGEM learned from all the data returns graphs that are reasonably close to the ground truth. To mimic the situation where we only have access to limited data, we sample a smaller dataset with K streams from the full dataset. We also sample $|S| = \{0, 5, 10, 20\}$ correct parental statements from the ground truth (0 statements implies no qualitative knowledge). We fix weight $w = 10$ and study the impact of the number of statements provided by the user and limited dataset size.

Figure 4 shows mean F1 scores for structure learning over 50 sampled subsets of 3 real-world datasets. We observe that the datasets vary in how qualitative statements aid the task, as gauged by the vertical distance between curves. Mimic has 75 labels and short event streams with a diverse set of labels, making it challenging to learn the structure even with additional soft knowledge. This is a case where hard constraints would improve performance, with the risk of overriding data given incorrect knowledge. On the other hand, LinkedIn has only 10 labels and shows knowledge having a substantial impact on structure discovery performance.

Parental Statements (Process Dependence): Expert Assessed Causal Pairs. We consider a real-world **Diabetes** dataset, with information about blood glucose levels, insulin dosage, eating and exercise routines of 70 diabetes patients (Frank and Asuncion 2010). We convert data for each patient into an event dataset, and use pairwise causal relation expert assessments from Acharya (2014) as qualitative knowledge in the form of parental statements. These include relations such as 'blood glucose measurement increase leads to less meal ingestion'. We study the effect of 11 such statements on structure discovery with limited data.

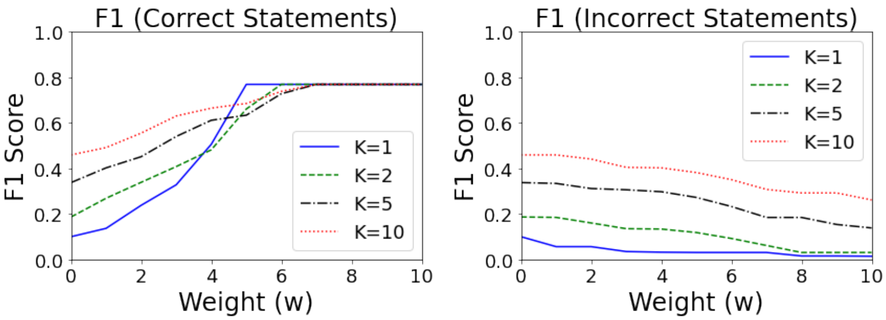

Figure 3: F1 score as a function of confidence (w) about parental statements for model #3 in Fig 1, shown for number of data streams $K \in \{1, 2, 5, 10\}$. Left: Correct statements. Right: Incorrect statements.

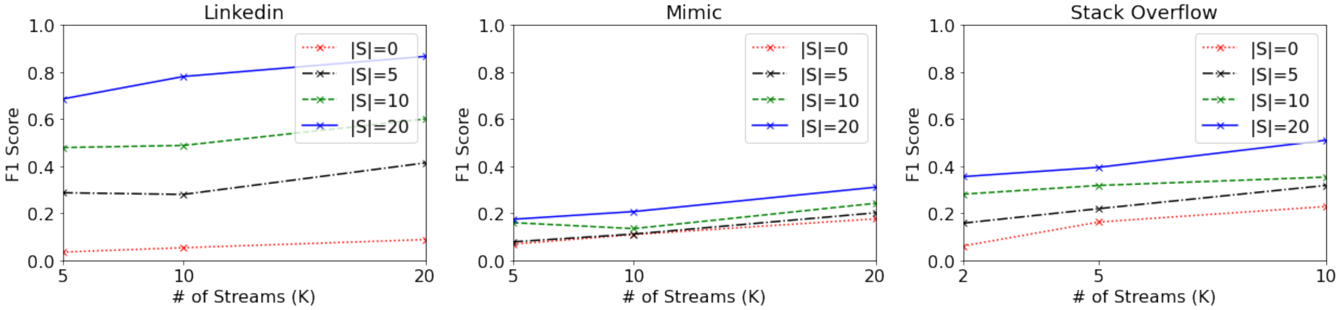

Figure 4: Comparing mean F1 scores from learning the structure of the temporal dynamics in 3 real world datasets using $K = \{5, 10, 20\}$ data streams for Linkedin and Mimic, and $K = \{2, 5, 10\}$ data streams for Stack Overflow. We also vary the number of correct parental statements sampled from the full dataset, $|S| = \{0, 5, 10, 20\}$.

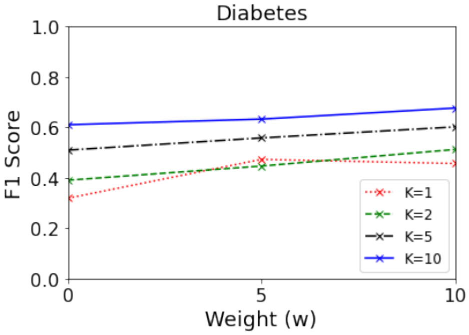

Figure 5: Mean F1 score from learning the structure in the Diabetes dataset using $K = \{1, 2, 5, 10\}$ data streams, potentially augmented with 11 pairwise causal relations.

Figure 5 shows mean F1 scores for structure learning over 50 sampled subsets of a varying number of event streams (patients), $K = \{1, 2, 5, 10\}$. On average, a stream includes $\sim$ 2 months of events for a patient. Weight $w = 0$ implies there is no qualitative knowledge. We see further evidence here that pairwise statements can help a score-based learner enhance its performance through such augmented knowledge when data is limited. As a point of comparison, note that if the structure is recovered using only the 11 pairwise statements, the F1 score is only 0.27 (for any K). This would be the case if the statements were written as temporal logical rules as in Li et al. (2020), illustrating the advantage of our knowledge augmentation approach.

6 Conclusions

We have pursued an important research question: how to learn GEMs in the low-data regime through incorporating background knowledge in different forms, seeking motivation from many real-world applications with limited data but access to high-level knowledge about event dynamics. For a broad class of GEMs that are expressible as piece-wise constant graphical event models, we proposed a novel Bayesian Gamma score including simplifying assumptions for making such prior assessments practically feasible. Simplifying such a score to 2 or 3 parameters however implies strong prior assumptions about process homogeneity.

In addition and complementary to *any score*, we therefore also consider other more practical forms of knowledge, proposing a general ‘incompatibility’ framework that allows augmenting limited data with qualitative background knowledge for score-based learning of GEMs. We presented techniques to incorporate qualitative statements pertaining to parameters for the special case of PGEMs, as well as more general structural statements made concrete via process (in)dependence. Potential future research directions include expanding the scope of forms/sources of knowledge and pursuing novel algorithms for knowledge augmentation.

Acknowledgments

We thank the anonymous reviewers for helpful feedback and suggestions. Xiao Shou was supported by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

References

Aalen, O. O.; Borgan, O.; and Gjessing, H. K. 2008. *Survival and Event History Analysis: A Process Point of View*. New York, NY, USA: Springer Science & Business Media.

Acharya, S. 2014. *Causal Modeling and Prediction over Event Streams*. Ph.D. thesis, University of Vermont.

Bhattacharjya, D.; Gao, T.; and Subramanian, D. 2020. Order-dependent event models for agent interactions. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1977–1983.

Bhattacharjya, D.; Subramanian, D.; and Gao, T. 2018. Proximal graphical event models. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 8147–8156.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov): 507–554.

Daley, D. J.; and Vere-Jones, D. 2002. *An Introduction to the Theory of Point Processes: Elementary Theory and Methods, volume I*. Springer, 2nd edition.

Dean, T.; and Kanazawa, K. 1989. A model for reasoning about persistence and causation. *Computational Intelligence*, 5: 142–150.

Didelez, V. 2008. Graphical models for marked point processes based on local independence. *Journal of Royal Statistical Society, Ser. B*, 70(1): 245–264.

Du, N.; Dai, H.; Trivedi, R.; Upadhyay, U.; Gomez-Rodriguez, M.; and Song, L. 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1555–1564. ACM.

Eichler, M. 1999. *Graphical Models in Time Series Analysis*. Ph.D. thesis, University of Heidelberg, Germany.

Eichler, M.; Dahlhaus, R.; and Dueck, J. 2017. Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2): 225–242.

Etesami, J.; Kiyavash, N.; Zhang, K.; and Singhal, K. 2016. Learning network of multivariate Hawkes processes: A time series approach. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 162–171.

Feelders, A. 2007. A new parameter learning method for Bayesian networks with qualitative influences. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 117–124.

Frank, A.; and Asuncion, A. 2010. UCI Machine Learning Repository.

Gao, T.; Subramanian, D.; Shanmugam, K.; Bhattacharjya, D.; and Mattei, N. 2020. A multi-channel neural graphical event model with negative evidence. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 3946–3953.

Grant, S.; and Betts, B. 2013. Encouraging user behaviour with achievements: An empirical study. In *Proceedings of the IEEE Working Conference on Mining Software Repositories (MSR)*, 65–68.

Gunawardana, A.; and Meek, C. 2016. Universal models of multivariate temporal point processes. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 556–563.

Gunawardana, A.; Meek, C.; and Xu, P. 2011. A model for temporal dependencies in event streams. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 1962–1970.

Hassanzadeh, O. 2021. Building a knowledge graph of events and consequences using Wikidata. In *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021)*.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1994. Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 293–301.

Heindorf, S.; Scholten, Y.; Wachsmuth, H.; Ngomo, A.-C. N.; and Potthast, M. 2020. CauseNet: Towards a causality graph extracted from the Web. In *International Conference on Information and Knowledge Management (CIKM)*.

Keiding, N. 1999. Event history analysis and inference from observational epidemiology. *Statistics in Medicine*, 18: 2353–2363.

Li, S.; Wang, L.; Zhang, R.; Chang, X.; Liu, X.; Xie, Y.; Qi, Y.; and Song, L. 2020. Temporal logic point processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 5990–6000.

Liao, W.; and Ji, Q. 2009. Learning Bayesian network parameters under incomplete data with domain knowledge. *Pattern Recognition*, 42(11): 3046–3056.

Lucas, P. J. F. 2005. Bayesian network modelling through qualitative patterns. *Artificial Intelligence*, 163(2): 233–263.

Luo, D.; Xu, H.; Zhen, Y.; Ning, X.; Zha, H.; Yang, X.; and Zhang, W. 2015. Multi-task multi-dimensional Hawkes processes for modeling event sequences. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 3685–3691.

Meek, C. 2014. Toward learning graphical and causal process models. In *Proceedings of Uncertainty in Artificial Intelligence (UAI) Workshop Causal Inference: Learning and Prediction*, 43–48.

Mei, H.; and Eisner, J. M. 2017. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems (NeurIPS)*, 6754–6764.

Mogensen, S. W.; Malinsky, D.; and Hansen, N. R. 2018. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 350–360.

Murphy, K. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California Berkeley, USA.

Omi, T.; Ueda, N.; and Aihara, K. 2019. Fully neural network based model for general temporal point processes. *arXiv preprint arXiv:1905.09690*.

Parikh, A. P.; Gunawardana, A.; and Meek, C. 2012. Conjoint modeling of temporal dependencies in event streams. In *Proceedings of Uncertainty in Artificial Intelligence (UAI) Bayesian Modelling Applications Workshop*.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 0-934613-73-7.

Saeed, M.; Villarroel, M.; Reisner, A. T.; Clifford, G.; Lehman, L.-W.; Moody, G.; Heldt, T.; Kyaw, T. H.; Moody, B.; and Mark, R. G. 2011. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39(5): 952.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 3027–3035.

Schweder, T. 1970. Composable Markov processes. *Journal of Applied Probability*, 7(2): 400–410.

Shchur, O.; Biloš, M.; and Günnemann, S. 2019. Intensity-free learning of temporal point processes. *arXiv preprint arXiv:1909.12127*.

Silander, T.; Kontkanen, P.; and Myllymaki, P. 2007. On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 360–367.

Wittig, F.; and Jameson, A. 2000. Exploiting qualitative knowledge in the learning of conditional probabilities of Bayesian networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Xiao, S.; Yan, J.; Yang, X.; Zha, H.; and Chu, S. M. 2017. Modeling the intensity function of point process via recurrent neural networks. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 17, 1597–1603.

Xu, H.; Luo, D.; and Zha, H. 2017. Learning Hawkes processes from short doubly-censored event sequences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 3831–3840.

Yu, X.; Shanmugam, K.; Bhattacharjya, D.; Gao, T.; Subramanian, D.; and Xue, L. 2020. Hawkesian graphical event models. In *Proceedings of the International Conference on Probabilistic Graphical Models (PGM)*.

Zhang, W.; Panum, T. K.; Jha, S.; Chalasani, P.; and Page, D. 2020. CAUSE: Learning Granger causality from event sequences using attribution methods. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Zhang, Y.; Sharma, K.; and Liu, Y. 2021. VigDet: Knowledge informed neural temporal point process for coordination detection on social media. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3218–3231.

Zhang, Y.; and Yan, J. 2021. Neural relation inference for multi-dimensional temporal point processes via message passing graph. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 3406–3412.

Zhou, K.; Zha, H.; and Song, L. 2013a. Learning social infectivity in sparse low-rank networks using multidimensional Hawkes processes. In *Artificial Intelligence and Statistics*, 641–649. PMLR.

Zhou, K.; Zha, H.; and Song, L. 2013b. Learning triggering kernels for multi-dimensional Hawkes processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1301–1309.

Zhou, Y.; Fenton, N.; and Neil, M. 2014. Bayesian network approach to multinomial parameter learning using data and expert judgments. *International Journal of Approximate Reasoning*, 55(5): 1252–1268.

Zuo, S.; Jiang, H.; Li, Z.; Zhao, T.; and Zha, H. 2020. Transformer Hawkes process. In *International Conference on Machine Learning*, 11692–11702.