

Dynamic Representation Learning with Temporal Point Processes for Higher-Order Interaction Forecasting

Tony Gracious, Ambedkar Dukkipati

Indian Institute of Science, Bangalore - 560012, INDIA
{tonygracious, ambedkar}@iisc.ac.in

Abstract

The explosion of digital information and the growing involvement of people in social networks led to enormous research activity to develop methods that can extract meaningful information from interaction data. Commonly, interactions are represented by edges in a network or a graph, which implicitly assumes that the interactions are pairwise and static. However, real-world interactions deviate from these assumptions: (i) interactions can be multi-way, involving more than two nodes or individuals (e.g., family relationships, protein interactions), and (ii) interactions can change over a period of time (e.g., change of opinions and friendship status). While pairwise interactions have been studied in a dynamic network setting and multi-way interactions have been studied using hypergraphs in static networks, there exists no method, at present, that can predict multi-way interactions or hyperedges in dynamic settings. Existing related methods cannot answer temporal queries like what type of interaction will occur next and when it will occur. This paper proposes a temporal point process model for hyperedge prediction to address these problems. Our proposed model uses dynamic representation learning techniques for nodes in a neural point process framework to forecast hyperedges. We present several experimental results and set benchmark results. As far as our knowledge, this is the first work that uses the temporal point process to forecast hyperedges in dynamic networks.

1 Introduction

Learning from temporal interactions between entities to extract meaningful information and knowledge is of paramount importance. For example, learning how a person interacts on social media can provide knowledge about that person’s preferences, and it can help in recommending items to that person. Similarly, an e-commerce website can better understand the users’ needs if it efficiently extracts knowledge from users’ consumption history. Previously these problems have been studied using representation or embedding learning in dynamic networks where interactions were modeled as instantaneous links or edges between two nodes formed at the time of interaction (Nguyen et al. 2018; Kumar, Zhang, and Leskovec 2019). For this, Temporal Point Processes (TPP) (Daley and Vere-Jones 2003)

have been introduced for modeling edge formation in dynamic networks. TPPs are stochastic processes that model localized events in time, and the events can be of multiple types (Hawkes 1971). To model dynamic networks, one represents each edge as an event type, and a probability distribution is defined over the time of its formation. Here, the probability distribution is parameterized using an intensity function based on representations of nodes. These node representations are functions of time and past interaction events. The parameters of these functions are learned by minimizing the negative log-likelihood of the interactions in the training data.

However, most real-world interactions are more complex than just pairwise interactions. For example, a person can have multiple items in a single shopping order, a group of people can co-author an article, mutual funds have stocks of various companies, and so on. These types of problems have been studied using hypergraphs, and many theoretical advances have taken place (Ghoshdastidar and Dukkipati 2017a,b).

A common technique employed to deal with this problem is to approximate these multiway interactions with pairwise interactions, which amounts to approximating hypergraphs with graphs. This leads to enormous information loss, as demonstrated in Figure 1. In this example, two different kinds of interaction between nodes $\{v_1, v_2, v_3, v_4, v_5, v_6\}$ have the same pairwise interaction graph. Further, it is impossible to infer the original interactions once they are projected into a pairwise graph.

This paper addresses the problem of forecasting higher-order interactions as hyperedge events using TPP. We define a conditional intensity function on each hyperedge that takes node representations as inputs. Since each hyperedge can have a variable number of nodes, we use a self-attention-based architecture for hyperedge encoding. Since nodes evolve as they interact, it is essential to employ dynamic node representations. In earlier works on dynamic networks (Trivedi et al. 2019; Dai et al. 2016; Cao et al. 2021), node embeddings are updated based on the node embedding of the other node in the interaction. For example, consider edge event (v_a, v_b) occurring at time t . To update node embeddings of v_a , we use the node embeddings of v_b and vice versa. However, hyperedge events have a variable number of nodes (Figure 1), so the techniques devel-

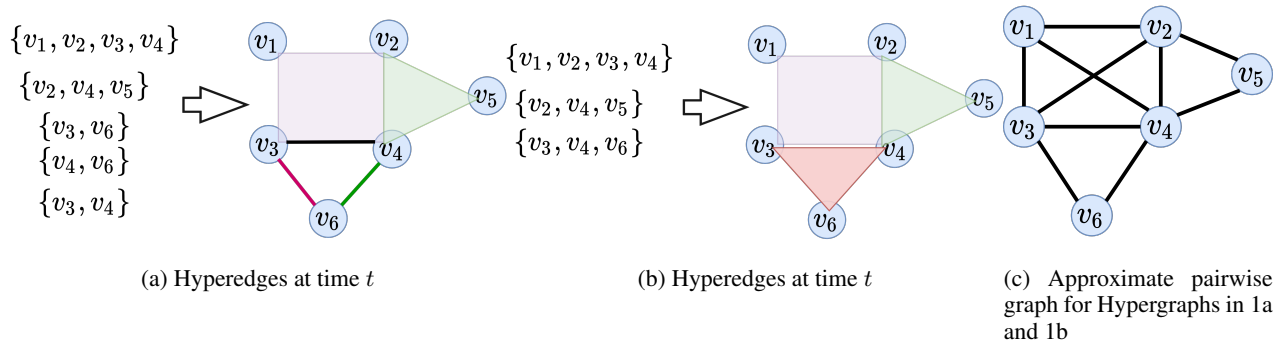


Figure 1: Higher-order interactions at time t are shown as hyperedges in Figures 1a and 1b. Here, hyperedges are represented by geometric shapes with their ends/corners showing the nodes and color showing their identity. One can see two different hypergraphs having the same projected graph in Figure 1c. This demonstrate a need for a technique that predicts hyperedges without approximating the hypergraph with a pairwise graph.

oped for pairwise interaction are not directly applicable to higher-order interaction. Hence, to address this, we use a self-attention-based encoding for node updates with parameters shared with the hyperlink prediction model.

Contributions. We propose a model called *Hypergraph Dynamic Hyperedge* (HGDHE) to model higher-order interactions as hyperedge events in a dynamic Hypergraph. Further, one can notice that these interactions will not always be between a homogeneous set of nodes. Hence, we also developed a bipartite hyperedge variant of our model called *Hypergraph Bipartite Dynamic Hyperedge* (HGBDHE). This will help in modeling interactions between two different types of nodes. Our contributions are as follows. (1) A temporal point process framework for hyperedge modeling that can forecast the type and time of interaction; (2) A model for representation learning for higher-order interaction data; (3) Extensive experiments on real-world datasets on both homogeneous and bipartite hyperedges; (4) Empirical results on performance gain obtained when we use hyperedges instead of pairwise modeling; and (5) Empirical results on performance gain are obtained when we use dynamic models instead of static models.

Related Works. Earlier works in modeling temporal information into networks can be categorized as (i) discrete-time models and (ii) continuous-time models. In discrete-time models, time is discretized into bins of equal size, and recurrent neural network-based models are used for modeling temporal evolution (Gupta, Sharma, and Dukkupati 2019; Gracious et al. 2021). Since discretization results in information loss and selecting bin size is a difficult task, the recent focus has been on continuous-time models (Kumar, Zhang, and Leskovec 2019; da Xu et al. 2020; Rossi et al. 2020). Unlike these discrete-time models, TPP-based continuous-time models (Shchur et al. 2021) can predict both dynamic interaction and time of interaction. Recently neural network-based TPP has been proposed to model dynamic interaction. However, these works approximate higher-order interactions with pairwise interactions (Dai

et al. 2016; Trivedi et al. 2019; Cao et al. 2021). It has been shown in Hyper-SAGNN (Zhang, Zou, and Ma 2019) that directly modeling the higher-order interaction will result in better performance than decomposing them into pairwise interactions.

Higher-order interaction between nodes can be modeled as link prediction in a hypergraph. Earlier works use matrix completion techniques to predict hyperedge. Coordinated Matrix Minimization (CMM) (Zhang et al. 2018), infer the missing hyperedges in the network by modeling adjacency matrix using a non-negative matrix factorization based latent representation. Recent works mainly concentrate on neural network-based scoring functions as they perform better than matrix completion-based techniques and are easier to train. Hyperpath (Huang, Liu, and Song 2019) model’s hyperedge as a tuple and uses a neural network-based scoring function to predict links. This method cannot model higher-order interactions as it expects the hyperedge size to be uniform for all edges. HyperSANN (Zhang, Zou, and Ma 2019) uses a self-attention-based architecture for predicting hyperlinks. It can learn node embeddings and predict non-uniform hyperlinks. However, it is a static model and cannot model the dynamic nature of hyperedges. The work presented in this paper is the first work that uses TPPs to forecast hyperedges when networks are evolving with time.

2 Dynamic Hyperedge Forecasting

Given a set of nodes $\mathcal{V} = \{v_1, v_2, \dots, v_{|\mathcal{V}|}\}$, an interaction event between a subset of these nodes is modeled as a hyperedge (h_i) with the time of interaction as its edge attribute. Here, $h_i \in \mathcal{H}$ is a subset of nodes in \mathcal{V} , and \mathcal{H} is a set of all valid combinations of nodes. Given the historical events $\mathcal{E}(t_m) = \{(h_1, t_1), \dots, (h_m, t_m)\}$ till time t_m , aim is to forecast future hyperedge h_i at time $t > t_m$.

2.1 Hyperedge Event Modeling

Given a hyperedge $h = \{v_1, v_2, \dots, v_k\}$, the probability of h occurring at time t can be modeled using TPP with condi-

tional intensity $\lambda_h(t)$ as

$$p_h(t) = \lambda_h(t)\mathcal{S}_h(t), \quad (1)$$

where $\mathcal{S}_h(t)$ is the survival function that denotes the probability that no event happened during the interval $[t_h^p, t]$ for hyperedge h . This is defined as

$$\mathcal{S}_h(t) = \exp\left(\int_{t_h^p}^t -\lambda_h(\tau)d\tau\right), \quad (2)$$

where $t_h^p = \max_{v \in h} t_v^p$ and t_v^p denotes the most recent interaction time of node v in h . We provide a more detailed background explanation of TPP in Appendix B in the supplementary material (Gracious and Dukkipati 2023). The conditional intensity function $\lambda_h(t)$ is parameterized by defining a positive function over the embeddings of nodes in h as,

$$\lambda_h(t) = f(\mathbf{v}_1(t), \mathbf{v}_2(t), \dots, \mathbf{v}_k(t)). \quad (3)$$

Here $f(\cdot) \geq 0$ can be realized by neural network for hyperedge events, and $\mathbf{v}_i(t) \in \mathbb{R}^d$ is the node embeddings at time t for node v_i . We follow the same architecture of the hyperedge modeling technique as Hyper-SAGNN (Zhang, Zou, and Ma 2019) with a final softplus layer as explained in Appendix C.1 in the supplementary material. In this paper, we have also developed baseline models using piece-wise constant node embeddings with intensity defined by Rayleigh Process (Ghosh 2009) to create an equivalent model for DeepCoevolve (Dai et al. 2016) in hyperedge interactions, as shown below,

$$\lambda_h(t) = f\left(\mathbf{v}_1(t_h^p), \mathbf{v}_2(t_h^p), \dots, \mathbf{v}_k(t_h^p)\right)(t - t_h^p). \quad (4)$$

This formulation will give us a closed-form solution for survival function, thereby for the probability of the event. Otherwise, integration has to be approximated by sampling.

2.2 Dynamic Node Representation

For each node in the network, we learn a low dimensional embedding $\mathbf{v}(t) \in \mathbb{R}^d$ that changes with time. It is done through three stages, i) Temporal drift, ii) History aggregation, and iii) Interaction update, as

$$\mathbf{v}(t) = \tanh\left(\underbrace{\mathbf{W}_0 \mathbf{v}(t_v^{p+})}_{\text{Interaction Update}} + \underbrace{\mathbf{W}_1 \Phi(t - t_v^p)}_{\text{Temporal Drift}} + \underbrace{\mathbf{W}_2 \mathbf{v}^s(t_{i-1})}_{\text{History Aggregation}} + \mathbf{b}_0\right). \quad (5)$$

Here, $\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d \times d}$, $\mathbf{b}_0 \in \mathbb{R}^d$ are learnable parameters, $\mathbf{v}(t_v^{p+})$ is the node embedding just after previous interaction for node v at time t_v^p and t_{i-1} is the last event time, $t_{i-1} < t$.

Temporal drift. This term models the inter-event evolution of a node with time. For a node v with previous event time t_v^p , the drift in embedding at time t is modelled by $\mathbf{W}_1 \Phi(t - t_v^p)$, where $\Phi(t) \in \mathbb{R}^d$ is Fourier time features (da Xu et al. 2020; Cao et al. 2021) defined as $\Phi(t) = [\cos(\omega_1 t + \theta_1), \dots, \cos(\omega_d t + \theta_d)]$. Here, $\{\omega_i\}_{i=1}^d$, and $\{\theta_i\}_{i=1}^d$ are learnable parameters.

History aggregation. This stage uses hypergraph convolution-based feature aggregation to incorporate the effect of past events. For this, we will construct a hypergraph using the past M events, $\{(h_{i-M}, t_{i-M}), \dots, (h_{i-1}, t_{i-1})\}$, and uses its incidence matrix $\mathbf{H}(t_{i-1}) \in \mathbb{R}^{|V| \times M}$ to apply hypergraph convolution as $[\mathbf{v}_1^s(t_{i-1}), \dots, \mathbf{v}_{|V|}^s(t_{i-1})] = \text{HGNN}\left(\mathbf{H}(t_{i-1}), [\mathbf{v}_1(t_{v_1}^{p+}), \dots, \mathbf{v}_1(t_{v_{|V|}}^{p+})]\right)$. Here, HGNN is a hypergraph graph convolution defined as in (Bai, Zhang, and Torr 2021).

Interaction update. When a node v is involved in an interaction h , it is influenced by the nodes it interacts within h . For extracting features of interaction, we use the dynamic embedding \mathbf{d}_v^h calculated as a function of embeddings of nodes $h - \{v\}$ at time t . The architecture of calculating this is shared with conditional intensity function as shown in Equation 13 in Appendix C.1 in the supplementary material. The entire update equation is $\mathbf{v}(t^+) = \tanh(\mathbf{W}_3 \mathbf{v}(t_v^{p+}) + \mathbf{W}_4 \Phi(t - t_v^p) + \mathbf{W}_5 \mathbf{d}_v^h + \mathbf{b}_1)$. Here, $\mathbf{W}_3, \mathbf{W}_4, \mathbf{W}_5 \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_1 \in \mathbb{R}^d$ are learnable parameters. If node v is involved in multiple hyperedge events $\{h_i, h_{i+1}, \dots, h_{i+L}\}$, then we will take the mean of dynamic embeddings from all of the hyperedges. Here, L is the number of concurrent hyperedges.

3 Learning Procedure

3.1 Loss Function

Once intensity parameterization is fixed for temporal point process as in Equation 3 or 4, the likelihood for hyperedge events $\mathcal{E}(T) = \{(h_1, t_1), \dots, (h_m, t_m)\}$ occurring in an interval $[0, T]$ can be modeled as, $p(\mathcal{E}(T)) = \prod_{i=1}^m p_{h_i}(t_i) \prod_{h \in \mathcal{H}} \mathcal{S}_h(t_h^\ell, T)$. Here, $p_{h_i}(t_i)$ is the probability of hyperedge event h_i occurring at time t_i as defined in Equation 1, $\mathcal{S}_h(t_h^\ell, T)$ is the probability that no event occurred for hyperedge h for the interval $[t_h^\ell, T]$, and t_h^ℓ is the last time occurrence for h ($t_h^\ell = 0$, when no event of h is observed). The loss for learning parameters can be found by taking the negative of log-likelihood as, $\mathcal{L} = -\sum_{i=1}^m \log(\lambda_{h_i}(t_i)) + \sum_{h \in \mathcal{H}} \int_0^T \lambda_h(t) dt$. Here, the first term corresponds to the sum of the negative log intensity of occurred events. The second term corresponds to the sum of intensities of all events. The following happens when we minimize the loss, the intensity rate of occurring events increases due to minimization of the first term, and intensity rates of events not occur decrease due to minimization of the second term. However, directly implementing this equation is computationally inefficient as $|\mathcal{H}| \leq 2^{|V|}$ is very large. Further, the integration in the second term does not always have a closed-form expression. In the next section, we will give this model a computationally efficient mini-batch training procedure.

3.2 Mini-Batch Loss

We divide the event sequences into independent segments to make backpropagation through time feasible, as done in previous works on pair-wise dynamic networks (Dai

et al. 2016). Then the loss for each segment is calculated as follows, for each (h_i, t_i) in the segment $\mathcal{E}_M = \{(h_1, t_1), \dots, (h_M, t_M)\}$, we use Monte-Carlo integration to find the log survival term of $p_{h_i}(t_i)$. Then the negative log-likelihood for that event is,

$$\mathcal{T}^s = \{t_j^s\}_{j=1}^N \leftarrow \text{Uniform}(t_{i-1}, t_i, N)$$

$$\mathcal{L}_{h_i} = -\log(\lambda_{h_i}(t_i)) + \sum_{j=2}^N (t_j^s - t_{j-1}^s) \lambda_{h_i}(t_j^s). \quad (6)$$

Here, \mathcal{T}^s is the set of uniformly sampled time points from the interval $[t_{i-1}, t_i]$. Then to consider the interactions events $h \in \mathcal{H}$ that were not observed during the above period, we sample some negative hyperedges for each interaction event (h_i, t_i) as described below,

1. Choose the size of negative hyperedge k based on a categorical distribution over hyperedge sizes observed in the training data. Here, parameters of the categorical distribution are learned from the training dataset.
2. Sample $\min(\lceil k/2 \rceil, |h_i|)$ nodes from the hyperedge h_i and rest of the nodes from $\mathcal{V} - h_i$. This strategy will avoid trivial negative samples.

Following the above steps, we sample $\mathcal{H}_i^n = \{h_1^n, \dots, h_B^n\}$ negative hyperedges, and for each of them, we calculate the negative log-likelihood for events not happening using Monte-Carlo integration. Then Equation 6 becomes,

$$\mathcal{L}_{h_i} = -\log(\lambda_{h_i}(t_i)) + \sum_{h \in \mathcal{H}_i^n \cup \{h_i\}} \sum_{j=2}^N (t_j^s - t_{j-1}^s) \lambda_h(t_j^s).$$

The final mini-batch loss is calculated by summing all \mathcal{L}_{h_i} for the events (h_i, t_i) in \mathcal{E}_M , $\sum_{i=1}^M \mathcal{L}_{h_i}$. Then the gradients are backpropagated for this loss, and the above training procedure is repeated in the next segment.

4 Extending to Bipartite Hyperedges

A higher-order interaction between nodes of two different type can be represented as a bipartite hyperedge $h = (\{v_1, \dots, v_k\}, \{v_{1'}, \dots, v_{k'}\})$. Here, $\{v_1, \dots, v_k\} \in \mathcal{H}$ is the left hyperedge with nodes from set \mathcal{V} , $\{v_{1'}, \dots, v_{k'}\} \in \mathcal{H}'$ is the right hyperedge with nodes from set \mathcal{V}' , and $\mathcal{V} \cap \mathcal{V}' = \emptyset$ (Sharma et al. 2021). For defining conditional intensity function, similar to homogenous hyperedges, we define $\lambda_h(t) = f(\{v_1, \dots, v_k\}, \{v_{1'}, \dots, v_{k'}\})$. Here $f(\cdot) \geq 0$ is defined by a neural network, and $\mathbf{v}_i(t), \mathbf{v}_{i'}(t)$ are node embeddings of the nodes in h . We adopt the same architecture of CATSETMAT (Sharma et al. 2021) for defining $f(\cdot)$ as explained in Appendix C.2 in the supplementary material.

To learn good dynamic representation for nodes, we follow the same architecture explained in Section 2.2. But, the left and right hyperedges have their own set of parameters for the Temporal drift, History aggregation, and Interaction update stages. Further, for dynamic embeddings in the Interaction update stage, nodes in the left hyperedge have it as a function of embeddings of

Datasets	$ \mathcal{V} $	$ \mathcal{V}' $	$ \mathcal{E}(T) $	$ \mathcal{H} $	$ \mathcal{H}' $
email-Enron	143	N/A	10,883	1,542	N/A
email-Eu	998	N/A	234,760	25,791	N/A
congress-bills	1,718	N/A	260,851	85,082	N/A
NDC-classes	1,161	N/A	49,724	1,222	N/A
NDC-sub	5,311	N/A	112,405	10,025	N/A
CastGenre	5,763	20	12,295	11,665	1,078
CastKeyword	3,998	1,953	13,826	12,365	10,737
CastCrew	5,763	4,541	12,295	11,665	9,451

Table 1: Datasets used for Homogeneous and Bipartite Hyperedges along with their vital statistics.

nodes in the right hyperedge, and vice versa. This function shares its parameters with the conditional intensity function.

For learning parameters, we follow the same procedure as that of homogenous hyperedges as explained in Section 3, except the negative sampling is done differently. We keep the left or right hyperedge fixed for generating negative samples and add a corrupted hyperedge on the other side. For example, for generating left hyperedge negative samples, we replace the right hyperedge by selecting a random subset of nodes from the right node-set \mathcal{V}' . The size of the corrupted hyperedge is selected based on the categorical distribution of sizes of the right hyperedge in the training set.

5 Experimental Settings

5.1 Datasets

All the datasets for homogeneous hyperedge interactions in Table 1 are taken from (Benson et al. 2018; Klimt and Yang 2004; Paranjape, Benson, and Leskovec 2017; Fowler 2006)¹. The bipartite hyperedges interactions are prepared from Kaggle’s The Movies Dataset ². A detailed explanation of all the datasets is given in Appendix D in the supplementary material. For homogeneous hypergraphs, $|\mathcal{V}'|$ is not applicable (N/A) as there is only one type of node, and $|\mathcal{H}'|$ is N/A as it is not directed.

5.2 Baselines

In Table 2, we have compared the properties of the baseline models we created against the proposed models HGDHE for the homogenous hyperedge interactions and HGBDHE for the bipartite hyperedge interactions. Here, models *Dynamic Edge* (DE) and *Dynamic Edge-drift* (DE-drift) use pairwise edge models instead of hyperedge interaction. Similarly, model *Bipartite Dynamic Edge* (BDE) uses bipartite pairwise edges to model bipartite hyperedge interaction. Further, in models *Rayleigh Hyperedge* (RHE) and *Rayleigh Dynamic Hyperedge* (RDHE), conditional intensity is modeled as the Rayleigh process as in Equation 4, and duration predictions are made using the closed-form expression in Equation 8. *Dynamic Hyperedge* (DHE) is the model that uses hyperedges for predicting higher-order interactions and has the

¹<https://www.cs.cornell.edu/arb/data/>

²<https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset>

Methods	T.D.	H.G.	I.U.	Hyperedge	Bipartite
RHE	✗	✗	✗	✓	✗
RDHE	✓	✗	✓	✓	✗
DE-drift	✓	✗	✗	✗	✗
DE	✓	✗	✓	✗	✗
DHE-drift	✓	✗	✗	✓	✗
DHE	✓	✗	✓	✓	✗
HGDHE-hist	✓	✓	✗	✓	✗
HGDHE	✓	✓	✓	✓	✗
BDE	✓	✗	✓	✗	✓
BDHE	✓	✗	✓	✓	✓
HGBDHE	✓	✓	✓	✓	✓

Table 2: Models and their properties. Here, ✓ indicates the usage of that property, and ✗ indicates the absence of that property. Here, T.D. corresponds to Temporal drift, H.G corresponds to History aggregation, and I.U. corresponds to Interaction update.

same dynamic node presentation as of DE model. So, in our studies, we will compare these two models to claim and establish the advantage of hyperedge modeling over pairwise modeling. Similarly, *Dynamic Hyperedge-drift* (DHE-drift) is the hyperedge model version of the DE-drift model. In the case of bipartite hyperedge interactions, we will compare models *Bipartite Dynamic Hyperedge* (BDHE) and BDE. A more detailed description of baselines can be found in Appendix E in the supplementary material.

Apart from the models mentioned above, we also provide comparisons against the current state-of-the-art approaches for temporal pairwise graphs DyRep (Trivedi et al. 2019), TGAT (da Xu et al. 2020), and TGN (Rossi et al. 2020). In this, DyRep uses the temporal point process framework for model edges and can predict both type and time of interaction. The other two models, TGAT and TGN, use a continuous time dynamic graph neural network for modeling edges and can only predict the type of interaction occurring at a particular time.

5.3 Prediction Tasks

Using temporal point process models, one can predict both the next event type and the time of the event. The following are the equivalent tasks in our settings.

Interaction type prediction. The type of interaction that occurs at time t can be predicted by finding the h_i with the maximum intensity value at that time as, $\hat{h} = \arg \max_{h_i} \lambda_{h_i}(t)$.

Interaction duration prediction. For interaction h occurred at time t_h^p , to predict the duration for future interaction, we have to calculate the expected time t with respect to the conditional distribution $p_h(t)$ in Equation 1,

$$\hat{t} = \int_{t_h^p}^{\infty} (t - t_h^p) p_h(t) dt. \quad (7)$$

If $\lambda(t)$ is modeled using a Rayleigh process as in Equation 4, we calculate the \hat{t} in close form as,

$$\hat{t} = \sqrt{\frac{\pi}{2 \exp f(\mathbf{v}_1(t_h^p), \mathbf{v}_2(t_h^p), \dots, \mathbf{v}_k(t_h^p))}}. \quad (8)$$

Otherwise, we have to compute the integration by sampling.

5.4 Metrics of Evaluation

We use Mean Reciprocal Rank (MRR) for evaluating the performance of interaction prediction at time t . To evaluate this, we find the reciprocal of rank (r_i) of the true hyperedge against candidate negative hyperedge in descending order of $\lambda_h(t)$ and then average them for all samples in the test set, $MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_{i+1}}$. Better performing models have higher MRR values. One can use Mean Absolute Error (MAE) for evaluating the performance of interaction duration prediction, $MAE = \frac{1}{N} \sum_{i=1}^N |\hat{t}_i - t_i^{true}|$. Here, better-performing models have lower MAE values.

5.5 Parameter Settings

For all experiments, we use the learning rate of 0.001, the embedding size d is fixed at 64 for homogeneous hyperedges and bipartite hyperedges, the batch size M is fixed as 128, and the negative sampling is fixed as $\mathcal{B} = 20$, and the training is done for 100 epochs. For the Monte Carlo estimate of log of survival probability in Section 3.2, we use $N = 20$ for datasets email-Enron, email-EU, and NDC-classes, $N = 5$ for NDC-sub and congress-bills datasets. The choice of N is made by considering memory constraints. All models are implemented PyTorch (Paszke et al. 2019), and all training is done using its Adam (Kingma and Ba 2015) optimizer. For all datasets, we use the first 50% of interactions for training, the next 25% for validation, and the rest for testing. The details of the computational infrastructure used are provided in the Appendix F in the supplementary material. All the reported scores are the average of ten randomized runs along with their standard deviation.

6 Results

6.1 Performance

In Table 3, one can see that the proposed model *Hypergraph Dynamic Hyperedge* (HGDHE) performs better than baselines in almost all the settings. Further, it significantly outperforms RHE, which uses static, and RDHE, which uses piece-wise constant node embeddings. Even though these models have a closed form expression for the event probability and duration estimation \hat{t} , their performance is poor compared to HGDHE, which has conditional intensity as a function of its dynamic node representation. We can also see that by comparing baselines, DHE to RDHE and RHE to DHE-drift, the models DHE and DHE-drift perform better as they use dynamic node representations as input to the conditional intensity function. This is because those models have more expressiveness and do not assume a parametric form for $\lambda_h(t)$.

Further, one can observe the advantage of using hyperedge for modeling higher-order interactions when comparing models DHE to DE and DHE-drift to DE-drift. This

Methods	email-Enron		email-Eu		congress-bills		NDC-classes		NDC-sub	
	MRR	MAE	MRR	MAE	MRR	MAE	MRR	MAE	MRR	MAE
DyRep ^a	42.0 ± 1.1	23.7 ± 1.6	32.2 ± 0.7	22.7 ± 0.8	40.3 ± 0.4	6.8 ± 0.1	65.8 ± 1.1	3.5 ± 0.2	55.7 ± 1.5	5.3 ± 0.4
TGAT ^b	52.6 ± 0.9	N/A	57.5 ± 0.8	N/A	78.4 ± 1.5	N/A	83.1 ± 0.7	N/A	77.0 ± 1.3	N/A
TGN ^c	54.3 ± 1.0	N/A	64.6 ± 1.2	N/A	80.6 ± 1.8	N/A	91.0 ± 0.8	N/A	79.3 ± 1.9	N/A
RHE	34.4 ± 0.8	127 ± 12	52.3 ± 0.4	26.4 ± 0.4	32.1 ± 0.3	54.1 ± 9.6	87.6 ± 1.6	8.9 ± 1.0	74.3 ± 0.3	4.6 ± 0.1
RDHE	26.7 ± 2.7	34.2 ± 0.4	27.6 ± 5.0	17.5 ± 0.6	52.9 ± 3.2	2.4 ± 0.2	81.1 ± 2.5	6.2 ± 0.9	66.4 ± 1.5	2.5 ± 0.0
DE-drift	30.8 ± 0.2	48.5 ± 0.6	43.4 ± 1.6	21.2 ± 0.0	43.2 ± 0.2	4.0 ± 0.3	60.6 ± 0.1	5.0 ± 0.1	64.9 ± 0.2	12.3 ± 0.8
DE	52.8 ± 0.3	16.5 ± 3.1	44.2 ± 0.7	19.3 ± 1.2	56.2 ± 2.8	2.6 ± 0.4	64.5 ± 0.7	3.6 ± 0.3	65.8 ± 1.4	13.8 ± 1.9
DHE-drift	50.2 ± 1.6	88.1 ± 4.5	58.3 ± 0.2	33.5 ± 0.8	84.5 ± 0.1	3.8 ± 0.2	88.6 ± 0.7	1.9 ± 0.1	79.3 ± 0.4	3.0 ± 0.0
DHE	60.5 ± 1.9	25.4 ± 4.3	64.0 ± 2.2	19.7 ± 2.0	92.2 ± 0.1	1.8 ± 0.2	88.9 ± 0.1	1.8 ± 0.2	86.5 ± 0.1	3.4 ± 0.1
HGDHE-hist	65.2 ± 1.2	18.2 ± 0.4	60.6 ± 0.0	24.6 ± 0.7	85.3 ± 0.1	3.4 ± 0.4	91.2 ± 0.7	1.4 ± 0.0	80.7 ± 0.1	1.7 ± 0.1
HGDHE	62.2 ± 2.8	16.1 ± 1.4	66.1 ± 2.9	15.1 ± 2.1	92.0 ± 0.0	1.6 ± 0.0	91.0 ± 0.3	1.2 ± 0.0	86.9 ± 0.5	1.6 ± 0.1

Table 3: Performance of dynamic homogeneous hyperedge forecasting in tasks of interaction type and interaction duration prediction. Here, interaction type prediction is evaluated using MRR in %, and interaction duration prediction is evaluated using MAE. The proposed model HGDHE beats baseline models in almost all the settings. **Citations:** ^a (Trivedi et al. 2019), ^b (da Xu et al. 2020), ^c (Rossi et al. 2020)

comparison is important because those models use the same dynamic node presentation, but DHE uses hyperedge modeling, and DE uses pairwise edge modeling. The same applies to the comparison of DHE-drift to DE-drift. Between DHE and DE, there is an improvement in the MRR metric in interaction type prediction tasks for all datasets. There is a 38.4% percentage gain in MRR for DHE compared to DE and a 40.3% gain for DHE-drift compared to DE-drift. In the interaction duration prediction task, we can observe a significant reduction in MAE for all the datasets except for the email-Enron and email-Eu datasets. A similar observation can be made between DHE-drift and DE-drift models. This is because more than 70% of interactions are pairwise in both of those datasets. Even though pairwise edges can achieve reasonable performance, we cannot identify the hyperedge among them if there are concurrent hyperedges with common nodes, as explained in Figure 1. Hence, hyperedge models perform better than pairwise models for interaction type prediction.

Comparison with previous works. In Table 3, we have compared our model HGDHE against previous works on pairwise temporal graphs DyRep, TGAT, and TGN. Here, we can see that our model considerably outperforms previous works as they are not hyperedge prediction models. The margin of performance is higher for datasets congress-bills and NDC-Sub compared to email-Enron and email-Eu as they have more higher-order interactions than the latter. Further, the low performance of DyRep compared to other works TGAT and TGN is because of the staleness in embedding (Kazemi et al. 2020) as it does not have a temporal drift stage to model evolve embedding during the interevent stage. The poor performance of TGAT and TGN is also because they use edge and node attributes along with graph neural networks to find the dynamic embeddings of the node. This information is not available for the datasets used in these experiments, so we initialized the node embed-

dings with random features and did not use edge features.

Performance on Bipartite interactions. In Table 4, one can observe that the proposed model *Hypergraph Bipartite Dynamic Hyperedge* (HGBDHE) performs better than all the other baselines in almost all settings and the models that use the bipartite property of the interaction perform better than their homogeneous counterparts. This can be inferred by the better performance of BDHE compared to DHE and HGBDHE compared to HGDHE. There is a gain of 23.6% in MRR and a reduction of 12.4% MAE for BDHE compared to DHE. Similarly, there is a gain of 36.2% in MRR and a reduction of 46.1% MAE for HGBDHE compared to HGDHE. This is because these models are more expressive and consider the bipartite nature of the interaction, as explained in Section 4. Similar to the case of homogeneous interactions, the pairwise model performs considerably poorer than the bipartite hyperedge models. It can be inferred from the poor performance of the BDE model, which uses a bipartite pairwise edge, compared to BDHE, which uses bipartite hyperedge for interaction modeling. Hence, one can conclude that bipartite hyperedge modeling represents data more accurately than pairwise modeling.

6.2 Ablation Studies

Role of interaction update. Here, we compare the performance of models that have this particular state in their update equation vs the models that do not. Firstly, one can observe a reduction of MAE by 24.7% and a gain of 3.9% in MRR for HGDHE compared to HGDHE-hist. Similarly, DHE outperforms DHE-drift in the interaction type prediction task for all datasets. One can also observe a significant reduction in MAE for email-Enron and email-Eu datasets. Between RHE and RDHE, RDHE has a performance gain in interaction duration prediction in all datasets. Even though RHE performs better in MRR in most datasets, MRR for RDHE is much better in congress-bills datasets. This pro-

Methods	CastGenre		CastKeyword		CastCrew	
	MRR	MAE	MRR	MAE	MRR	MAE
BDE	23.5 ± 0.7	10.3 ± 0.3	13.6 ± 0.1	21.9 ± 0.6	13.6 ± 0.2	22.2 ± 1.0
DHE	27.5 ± 0.8	2.8 ± 0.4	36.1 ± 1.3	15.3 ± 1.7	26.0 ± 1.7	9.2 ± 2.0
BDHE	33.2 ± 0.5	2.9 ± 0.2	38.7 ± 1.6	9.6 ± 2.3	37.2 ± 2.6	9.1 ± 1.3
HGDHE	27.5 ± 1.6	19.3 ± 1.6	35.3 ± 1.9	22.6 ± 1.5	25.1 ± 2.8	8.8 ± 2.0
HGBDHE	33.6 ± 1.5	4.1 ± 0.8	41.3 ± 1.7	9.2 ± 1.6	42.7 ± 2.0	8.7 ± 1.6

Table 4: Performance of dynamic bipartite hyperedge forecasting in tasks of interaction type and interaction duration prediction. Proposed model HGBDHE beats baseline models in almost in all the settings.

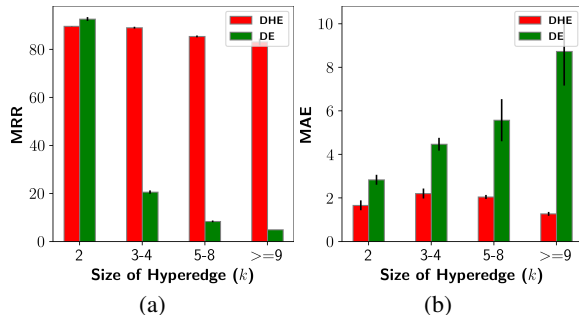


Figure 2: Figure 2a shows the effect of hyperedge size (k) on interaction type prediction, and Figure 2b shows the effect on interaction duration prediction for NDC-classes dataset

vides evidence that using Interaction update stage results in performance improvement.

Role of temporal drift. The advantage of this stage can be observed by comparing the performance of RHE to DHE-drift and RDHE to DHE. There is an average 44% gain in MRR and 42.03% reduction in MAE for DHE-drift compared to RHE. Further, DHE uniformly outperforms RDHE for interaction type prediction in all datasets. There is an average of 90% improvement in the MRR for the interaction type prediction task and a 13.7% decrease in the MAE error for the duration prediction task. Hence, Temporal drift stage helps in performance improvement.

Role of history aggregation. The advantage of this stage can be observed by comparing HGDHE to DHE and HGDHE-hist to DHE-drift. HGDHE outperforms DHE in all the tasks except in Congress bills. There is an average gain of 1.7% in MRR and a 31.75% reduction in MAE for HGDHE compared to DHE and HGDHE-hist outperforms DHE-drift in all settings. A similar observation can be made in bipartite datasets when comparing models HGBDHE and BDHE. Both models give a comparable performance for interaction duration prediction except for CastGenre dataset. For interaction type prediction HGBDHE achieves a gain of 7% in MRR when compared to BDHE.

6.3 Hyperedge Size (k) Vs Performance

While DHE and DE models use the same dynamic representations, they have different conditional intensity functions. DHE uses a hyperedge-based conditional intensity function, and DE uses a pairwise edge-based function. To compare them, we divide hyperedges into different groups based on their sizes, and the mean value of the evaluation metric is calculated for each group. These groups are defined so that each group has enough samples to make the comparisons statistically significant. In Figures 2a and 2b, we have shown the effect of hyperedge size on our model DHE, which uses hyperedge modeling, and compared it against model DE, which uses pairwise edges. Here, each hyperedge is grouped into groups with $k = 2$, $3 \leq k \leq 4$, $5 \leq k \leq 8$ and $k \geq 9$. From Figure 2a, we can see that performance of DE is poor for hyperedges with a size of more than 2, but our model DHE has almost similar performance for hyperedges of different sizes. The reason for this is that DHE is suited more for hyperedges of varying sizes compared to DE. We observed a similar trend in other datasets, as shown in Appendix G in the supplementary material (Gracious and Dukkipati 2023). For the interaction duration prediction task, the error for the DE model increases with hyperedge size while the DHE model performs similarly for all hyperedge sizes.

7 Concluding Remarks

Forecasting higher-order interactions in time-evolving hypergraphs is very challenging, and this has not been studied before. In this work, we provide the first solution to this problem by proposing a model for forecasting higher-order interaction between nodes in a network as temporal hyperedge formation events. For this, we develop a mechanism that uses the temporal point process to learn the dynamic representation of nodes. Our future work includes extending the proposed model for modeling multi-relational higher orders interactions (Fatemi et al. 2020) and incorporating multi-hop information into node representation using hypergraph neural network-based techniques for better predictive performance. We also plan to explore options to reduce the training time by using t -batch (Kumar, Zhang, and Leskovec 2019) based approach that can use parallel training techniques.

Acknowledgements

The authors would like to thank the SERB, Department of Science and Technology, Government of India, for the generous funding towards this work through the IMPRINT Project: IMP/2019/000383.

References

- Bai, S.; Zhang, F.; and Torr, P. H. 2021. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110: 107637.
- Benson, A. R.; Abebe, R.; Schaub, M. T.; Jadbabaie, A.; and Kleinberg, J. 2018. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*.
- Cao, J.; Lin, X.; Cong, X.; Guo, S.; Tang, H.; Liu, T.; and Wang, B. 2021. Deep Structural Point Process for Learning Temporal Interaction Networks. In Oliver, N.; Pérez-Cruz, F.; Kramer, S.; Read, J.; and Lozano, J. A., eds., *Machine Learning and Knowledge Discovery in Databases. Research Track*, 305–320.
- da Xu; chuanwei ruan; evren korpeoglu; sushant kumar; and kannan achan. 2020. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*.
- Dai, H.; Wang, Y.; Trivedi, R.; and Song, L. 2016. Deep coevolutionary network: Embedding user and item features for recommendation. arXiv:1609.03675.
- Daley, D. J.; and Vere-Jones, D. 2003. *An introduction to the theory of point processes. Vol. I*. Springer-Verlag.
- Fatemi, B.; Taslakian, P.; Vazquez, D.; and Poole, D. 2020. Knowledge Hypergraphs: Prediction Beyond Binary Relations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 2191–2197.
- Fowler, J. H. 2006. Legislative cosponsorship networks in the US House and Senate. *Social Networks*, 28(4): 454–465.
- Ghosh, J. K. 2009. Survival and Event History Analysis: A Process Point of View by Odd O. Aalen, Ørnulf Borgan, Håkon K. Gjessing. *International Statistical Review*, 77(3): 463–464.
- Ghoshdastidar, D.; and Dukkipati, A. 2017a. Consistency of Spectral Hypergraph Partitioning under Planted Partition Model. *The Annals of Statistics*, 45(1): 289–315.
- Ghoshdastidar, D.; and Dukkipati, A. 2017b. Uniform Hypergraph Partitioning: Provable Tensor Methods and Sampling Techniques. *The Journal of Machine Learning Research*, 18(50): 1–41.
- Gracious, T.; and Dukkipati, A. 2023. Supplementary Material for Dynamic Representation Learning with Temporal Point Processes for Higher-Order Interaction Forecasting. arXiv:2112.10154.
- Gracious, T.; Gupta, S.; Kanthali, A.; Castro, R. M.; and Dukkipati, A. 2021. Neural Latent Space Model for Dynamic Networks and Temporal Knowledge Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4054–4062.
- Gupta, S.; Sharma, G.; and Dukkipati, A. 2019. A Generative Model for Dynamic Networks with Applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7842–7849.
- Hawkes, A. G. 1971. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1): 83–90.
- Huang, J.; Liu, X.; and Song, Y. 2019. Hyper-Path-Based Representation Learning for Hyper-Networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19*, 449–458.
- Kazemi, S. M.; Goel, R.; Jain, K.; Kobayez, I.; Sethi, A.; Forsyth, P.; and Poupart, P. 2020. Representation learning for dynamic graphs: A survey. *J. Mach. Learn. Res.*, 21(70): 1–73.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.
- Klimt, B.; and Yang, Y. 2004. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, 217–226.
- Kumar, S.; Zhang, X.; and Leskovec, J. 2019. Predicting Dynamic Embedding Trajectory in Temporal Interaction Networks. In *Proceedings of the 25th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.
- Nguyen, G. H.; Lee, J. B.; Rossi, R. A.; Ahmed, N. K.; Koh, E.; and Kim, S. 2018. Continuous-time dynamic network embeddings. In *Companion Proceedings of the The Web Conference 2018*, 969–976.
- Paranjape, A.; Benson, A. R.; and Leskovec, J. 2017. Motifs in temporal networks. In *Proceedings of the tenth ACM international conference on web search and data mining*, 601–610.
- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32.
- Rossi, E.; Chamberlain, B.; Frasca, F.; Eynard, D.; Monti, F.; and Bronstein, M. 2020. Temporal graph networks for deep learning on dynamic graphs. arXiv:2006.10637.
- Sharma, G.; Singh, S. P.; Devi, V. S.; and Murty, M. N. 2021. The CAT SET on the MAT: Cross Attention for Set Matching in Bipartite Hypergraphs. arXiv:2111.00243.
- Shchur, O.; Türkmen, A. C.; Januschowski, T.; and Günnemann, S. 2021. Neural Temporal Point Processes: A Review. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4585–4593. International Joint Conferences on Artificial Intelligence Organization.
- Trivedi, R.; Farajtabar, M.; Biswal, P.; and Zha, H. 2019. DyRep: Learning Representations over Dynamic Graphs. In *International Conference on Learning Representations*.

Zhang, M.; Cui, Z.; Jiang, S.; and Chen, Y. 2018. Beyond Link Prediction: Predicting Hyperlinks in Adjacency Space. In *AAAI*.

Zhang, R.; Zou, Y.; and Ma, J. 2019. Hyper-SAGNN: a self-attention based graph neural network for hypergraphs. In *International Conference on Learning Representations*.