

# 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. 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 there have been 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  $\{v_1, v_2, v_3, v_4, v_5\}$  have the same pairwise interaction graph. Further, it is impossible to infer the original interactions once they are projected into a pairwise graph.

In this paper, we address 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 developed for pairwise interaction are not directly applicable to higher-order interaction. Hence, to address this, we use a

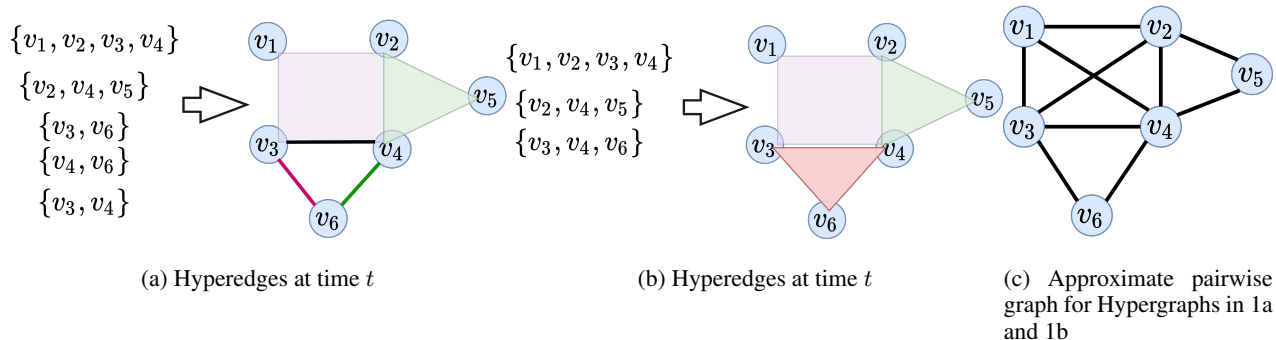


Figure 1: Higher-order interactions at time  $t$  are shown as hyperedges in Figures 1a and 1b. Here, hyperedges are represented by geometric shapes with their ends/corners showing the nodes and color showing their identity. One can see two different hypergraphs having the same projected graph in Figure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

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 Dukkipati 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 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 HyperSAGNN (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 set of all valid combination 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 conditional intensity  $\lambda_h(t)$  as

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

where  $\mathcal{S}_h(t)$  is survival function that denotes 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 in-teraction time of node  $v$  in  $h$ . We provide a more detailed background explanation of TPP in Appendix B. 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 hyper-edge 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 hyper-edge modeling technique as Hyper-SAGNN (Zhang, Zou, and Ma 2019) with a final softplus layer as explained in Appendix C.1. In this paper, we have also developed baseline models using piece-wise constant node embeddings with intensity defined by Rayleigh Process 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  $[\mathbf{v}_1^s(t_{i-1}), \dots, \mathbf{v}_{|V|}^s(t_{i-1})] = \text{HGNN}\left(\mathbf{H}(t_{i-1}), [\mathbf{v}_1(t_{i-1}^p), \dots, \mathbf{v}_{|V|}(t_{i-1}^p)]\right)$ . 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. 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  $t_h^\ell$  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  $\mathcal{V} \cap \mathcal{V}' = \emptyset$ . More details on these types of higher-order heterogeneous graphs can be found in the recent work (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 follow the same architecture of CAT-SETMAT (Sharma et al. 2021) for defining  $f(\cdot)$  as explained in Appendix C.2.

Now 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 nodes in the right hyperedge, and vice

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	2,065
CastKeyword	12,180	3,997	3,360	3,360	3,345
CastCrew	12,634	12,704	3,517	3,507	3,517

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

versa. This function shares its parameters with the conditional intensity function, as shown in Equation 16 in Appendix C.2.

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 work (Benson et al. 2018)<sup>1</sup>. The bipartite hyperedges interactions are prepared from Kaggle’s The Movies Dataset <sup>2</sup>. A detailed explanation of all the datasets is given in Appendix D. 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 9. *Dynamic Hyperedge* (DHE) is the model that uses hyperedges for predicting higher-order interactions and has the same dynamic node presentation as of DE model. So, in our

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

<sup>2</sup><https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata>

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.

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.

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 shown below,

$$\hat{h} = \arg \max_{h_i} \lambda_{h_i}(t). \quad (7)$$

**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,

$$\hat{t} = \int_{t_h^p}^{\infty} (t - t_h^p) p_h(t) dt. \quad (8)$$

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

$$\hat{t} = \sqrt{\frac{\pi}{2 \exp f(\mathbf{v}_1(t_h^p), \mathbf{v}_2(t_h^p), \dots, \mathbf{v}_k(t_h^p))}}. \quad (9)$$

Otherwise, we have to compute the integration by sampling.

### 5.4 Metrics of Evaluation

**Mean Average Reciprocal Rank (MRR).** We use this for evaluating the performance of interaction prediction at time  $t$ . For finding 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,  $MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_i+1}$ . Here, better-performing models have higher MRR values.

**Mean Average Error (MAE).** We use this 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. 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  $\lambda_h(t)$ .

Methods	email-Enron		email-Eu		congress-bills		NDC-classes		NDC-sub	
	MRR	MAE	MRR	MAE	MRR	MAE	MRR	MAE	MRR	MAE
DyRep <sup>a</sup>	42.08 ± 1.14	23.74 ± 1.68	32.23 ± 0.70	22.77 ± 0.87	40.30 ± 0.40	6.83 ± 0.12	65.80 ± 1.14	3.50 ± 0.20	55.71 ± 1.56	5.35 ± 0.43
TGAT <sup>b</sup>	52.60 ± 0.95	N/A	57.53 ± 0.89	N/A	78.47 ± 1.52	N/A	83.1 ± 0.72	N/A	77.06 ± 1.30	N/A
TGN <sup>c</sup>	54.32 ± 1.02	N/A	64.63 ± 1.2	N/A	80.61 ± 1.8	N/A	91.06 ± 0.81	N/A	79.33 ± 1.90	N/A
RHE	34.45 ± 0.81	127.19 ± 12.17	52.37 ± 0.47	26.43 ± 0.47	32.14 ± 0.38	54.13 ± 9.60	87.64 ± 1.63	8.98 ± 1.09	74.33 ± 0.30	4.66 ± 0.15
RDHE	26.73 ± 2.76	34.22 ± 0.49	27.68 ± 5.02	17.54 ± 0.68	52.90 ± 3.24	2.44 ± 0.22	81.17 ± 2.56	6.29 ± 0.91	66.46 ± 1.50	2.52 ± 0.07
DE-drift	30.84 ± 0.29	48.50 ± 0.65	43.47 ± 1.69	21.21 ± 0.06	43.27 ± 0.23	4.07 ± 0.32	60.64 ± 0.19	5.02 ± 0.15	64.90 ± 0.29	12.38 ± 0.84
DE	52.89 ± 0.38	16.52 ± 3.14	44.27 ± 0.74	19.37 ± 1.27	56.27 ± 2.89	2.65 ± 0.42	64.58 ± 0.78	3.60 ± 0.38	65.83 ± 1.41	13.82 ± 1.93
DHE-drift	50.25 ± 1.65	88.14 ± 4.59	58.38 ± 0.20	33.56 ± 0.82	84.50 ± 0.13	3.84 ± 0.26	88.68 ± 0.71	1.92 ± 0.12	79.31 ± 0.41	3.03 ± 0.01
DHE	60.57 ± 1.97	25.48 ± 4.37	64.03 ± 2.22	19.72 ± 2.00	<b>92.21 ± 0.19</b>	1.87 ± 0.22	88.93 ± 0.16	1.84 ± 0.21	86.52 ± 0.18	3.49 ± 0.14
HGDHE-hist	<b>65.26 ± 1.24</b>	18.25 ± 0.43	60.69 ± 0.09	24.66 ± 0.77	85.31 ± 0.10	3.44 ± 0.34	<b>91.24 ± 0.74</b>	1.42 ± 0.08	80.73 ± 0.15	1.75 ± 0.19
HGDHE	62.21 ± 2.85	<b>16.12 ± 1.45</b>	<b>66.12 ± 2.90</b>	<b>15.18 ± 2.14</b>	92.09 ± 0.03	<b>1.65 ± 0.06</b>	91.01 ± 0.35	<b>1.21 ± 0.04</b>	<b>86.92 ± 0.51</b>	<b>1.65 ± 0.18</b>

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:** <sup>a</sup> (Trivedi et al. 2019), <sup>b</sup> (da Xu et al. 2020), <sup>c</sup> (Rossi et al. 2020)

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 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 embeddings 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. We can see that 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

**Effect of interaction update on performance.** Here, we compare the performance of models that have this particular state in their update equation vs the models that do not. Firstly, we 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 considerably 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, one can see that RDHE has a performance gain in interaction duration prediction in all datasets. Even though RHE performs better in MRR in most datasets, one can see MRR for RDHE is much better in congress-bills datasets. Hence one can conclude that using `Interaction` update stage resulted in performance improvement.

**Effect of temporal drift on performance.** Similar to the earlier study, the advantage of this stage can be observed by comparing the performance of RHE to DHE-drift and

Methods	CastGenre		CastKeyword		CastCrew	
	MRR	MAE	MRR	MAE	MRR	MAE
BDE	23.55 ± 0.75	10.34 ± 0.34	13.61 ± 0.14	21.98 ± 0.60	13.61 ± 0.24	22.23 ± 1.03
DHE	27.58 ± 0.88	<b>2.88 ± 0.43</b>	36.18 ± 1.34	15.32 ± 1.75	26.03 ± 1.75	9.29 ± 2.01
BDHE	33.22 ± 0.52	2.91 ± 0.26	38.77 ± 1.69	9.61 ± 2.33	37.29 ± 2.65	9.18 ± 1.37
HGDHE	27.59 ± 1.60	19.39 ± 1.62	35.32 ± 1.96	22.64 ± 1.50	25.19 ± 2.83	8.85 ± 2.07
HGBDHE	<b>33.65 ± 1.58</b>	4.16 ± 0.84	<b>41.32 ± 1.74</b>	<b>9.27 ± 1.67</b>	<b>42.77 ± 2.00</b>	<b>8.77 ± 1.68</b>

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.

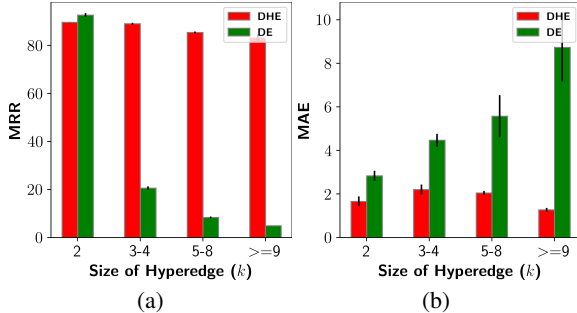


Figure 2: Figure 2a shows the effect of hyperedge size ( $k$ ) on interaction type prediction, and Figure 2b shows the effect on interaction duration prediction for NDC-classes dataset

RDHE to DHE. There is an average of 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.

**Effect of history aggregation on performance.** The advantage of this stage can be observed by comparing HGDHE to DHE and HGDHE-hist to DHE-drift. We can see 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. We can see 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 both the models DHE and DE use dynamic representations, they have different conditional intensity functions. While 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. 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

The problem of 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.; 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.; Kobzyev, 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*.

## A Notations

Table 5 shows the notations and their respective definitions used in this paper. We used bold lower-case letters to show vectors and bold upper-case letters to show matrices.

## B Background on Temporal Point Process

A TPP (Daley and Vere-Jones 2003) is a continuous-time stochastic process that models discrete events in time  $t_i$ ,  $t_i \in \mathbb{R}^+$ . It defines a conditional density function (CDF) for future event time  $t$  by observing historical event times till time  $t_n$ ,  $\mathcal{T}(t_n) = \{t_1, t_2, \dots, t_n\}$ . A convenient and interpretable way to parameterize CDF is by defining a conditional intensity function or instantaneous stochastic rate of events  $\lambda(t)$  defined below,

$$\lambda(t)dt = p(\text{event in } [t, t + dt] | \mathcal{T}(t)). \quad (10)$$

That is,  $\lambda(t)dt$  is the probability of observing an event in interval  $[t, t + dt]$  by observing history till  $t$ . Then we can write the CDF of the next event time as,

$$p(t | \mathcal{T}(t_n)) = \lambda(t)S(t),$$

$$S(t) = \exp\left(\int_{t_n}^t -\lambda(\tau)d\tau\right). \quad (11)$$

Here,  $S(t)$  is the probability that no event happened during the interval  $[t_n, t]$ , which is called the survival function.

The choice of the functional form of  $\lambda(t)$  depends upon the nature of the problem we are trying to model. For example, the Poisson process has constant intensity  $\lambda(t) = \mu$ . Hawkes process (Hawkes 1971) is used when events have self exciting nature with  $\lambda(t) = \mu + \alpha \sum_{t_i \in \mathcal{T}(t)} \mathcal{K}(t - t_i)$ . Here,  $\mu \geq 0$  is the base rate,  $\mathcal{K}(t) \geq 0$  is the excitation kernel, and  $\alpha \geq 0$  is the strength of excitation. Rayleigh process (Ghosh 2009) with  $\lambda(t) = \alpha t$ , where rate of events increase with time and  $\alpha > 0$  is the weight parameter. If the intensity is the output of neural network  $f(\cdot)$  that takes history as input,  $\lambda(t) = f(\mathcal{T}(t))$ , it is called Neural Temporal Point process (Shchur et al. 2021).

## C Conditional Intensity Function Architecture

### C.1 Homogeneous

Given the node embeddings of hyperedge  $h = \{v_1, v_2, \dots, v_k\}$ , the importance weights for each node are calculated by a self-attention layer, as given by,

$$e_{ij} = (\mathbf{W}_Q^T \mathbf{v}_i(t))^T \mathbf{W}_K^T \mathbf{v}_j(t), \forall 1 \leq i, j \leq k, i \neq j,$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{1 \leq \ell \leq k, i \neq \ell} \exp(e_{i\ell})}. \quad (12)$$

These weights are used to calculate the dynamic embeddings for each node  $v_i$  as,

$$\mathbf{d}_i^h = \tanh\left(\sum_{1 \leq j \leq k, i \neq j} \alpha_{ij} \mathbf{W}_V^T \mathbf{v}_j(t)\right). \quad (13)$$

Here,  $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$  are learnable weights. We also create static embeddings  $\mathbf{s}_i^h = \mathbf{W}_s \mathbf{v}_i(t)$ , where  $\mathbf{W}_s \in \mathbb{R}^{d \times d}$  is a learnable parameter. Then we calculate the Hadamard power of the difference between static and dynamic embedding pairs followed by a linear layer and average pooling to get the final score  $\mathcal{P}^h$  as shown below,

$$o_i = \mathbf{W}_o^T (\mathbf{d}_i^h - \mathbf{s}_i^h)^2 + b_o,$$

$$\mathcal{P}^h = \frac{1}{k} \sum_{i=1}^k \mathcal{P}_i^h = \frac{1}{k} \sum_{i=1}^k \log(1 + \exp(o_i)). \quad (14)$$

Here,  $\mathbf{W}_o \in \mathbb{R}^{d \times 1}, b_o \in \mathbb{R}$  are learnable parameters of output layer. In our model the value of  $f(\mathbf{v}_1(t), \mathbf{v}_2(t), \dots, \mathbf{v}_k(t)) = \mathcal{P}^h$ .

### C.2 Bipartite

For a bipartite hyperedge  $h$ , there are two sets of nodes with a directed relation between them. This can be represented as  $h = (\{v_1, v_2, \dots, v_k\}, \{v_1', v_2', \dots, v_{k'}'\})$ . Here, we use the cross attention between the sets of nodes to create dynamic embeddings, as shown below,

$$e_{ij'} = (\mathbf{W}_Q^T \mathbf{v}_i(t))^T \mathbf{W}_{K'}^T \mathbf{v}_{j'}(t), \forall 1 \leq i \leq k, 1 \leq j' \leq k',$$

$$e_{i'j} = (\mathbf{W}_Q^T \mathbf{v}_{i'}(t))^T \mathbf{W}_K^T \mathbf{v}_j(t), \forall 1 \leq i' \leq k', 1 \leq j \leq k$$

$$\alpha_{ij'} = \frac{\exp(e_{ij'})}{\sum_{1 \leq \ell' \leq k'} \exp(e_{i\ell'})}$$

$$\alpha_{i'j} = \frac{\exp(e_{i'j})}{\sum_{1 \leq \ell \leq k} \exp(e_{i'\ell})}. \quad (15)$$

These weights are used to find the dynamic embeddings for left node  $v_i$  and right node  $v_{i'}$  as follows,

$$\mathbf{d}_i^h = \tanh\left(\sum_{1 \leq j' \leq k'} \alpha_{ij'} \mathbf{W}_V^T \mathbf{v}_{j'}(t)\right)$$

$$\mathbf{d}_{i'}^h = \tanh\left(\sum_{1 \leq j \leq k} \alpha_{i'j} \mathbf{W}_V^T \mathbf{v}_j(t)\right). \quad (16)$$

Here,  $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$  are the learnable weights for nodes in the left, and  $\mathbf{W}_{Q'}, \mathbf{W}_{K'}, \mathbf{W}_{V'} \in \mathbb{R}^{d \times d}$  are the learnable weights for nodes in the right. Then static embeddings are created for both left  $\mathbf{s}_i^h = \mathbf{W}_s \mathbf{v}_i(t)$  and right side  $\mathbf{s}_{i'}^h = \mathbf{W}_s' \mathbf{v}_{i'}(t)$ , where  $\mathbf{W}_s, \mathbf{W}_s' \in \mathbb{R}^{d \times d}$  are learnable parameters. Then we calculate the Hadamard power of the difference between static and dynamic embedding pairs for both the left and right nodes, followed by a linear layer and average pooling to get final score  $\mathcal{P}^h$  as shown below,

Datasets	Definitions
$\lambda(t)$	Conditional intensity function
$t$	Time
$k$	Hyperedge size
$t^p$	Previous time
$dt$	Derivative of time $t$
$\mathbb{R}$	Real space with dimension 1
$\mathcal{V}$	Set of nodes
$\mathcal{T}$	Historical event times
$d$	Embedding size
$v_i$	$i$ th node
$h_i$	$i$ th hyperedge/higher-order interaction
$\mathcal{H}$	Set of all valid interactions
$\mathcal{E}$	Sequence of temporal ordered interaction events
$\lambda_h(t)$	Conditional intensity function of interaction $h$
$t_h^p$	Previous time of interaction for hyperedge $h$
$t^s$	Monte-carlo $t$ sample
$p(t)/p_h(t)$	Probability of the event/interaction $h$
$S(t)$	Survival function for the event
$t_v^p$	Previous time of interaction for node $v$
$t_h^l$	Last time occurrence of $h$
$f(\cdot)$	Scoring function for conditional intensity
$\theta$	Learnable parameter
$v(t)$	Node embedding at time $t$
$\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_0$	Learnable parameters for dynamic node embeddings
$\mathbf{W}_3, \mathbf{W}_4, \mathbf{W}_5, \mathbf{b}_1$	Learnable parameters for interaction update
$v^s(t)$	History aggregation embeddings
$\Phi(t)$	Time embedding function
$\{\omega_i\}_{i=1}^d$ and $\{\theta_i\}_{i=1}^d$	Learnable parameters of $\Phi(t)$
$\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$	Self-Attention learnable weights
$e_{ij}$	Attention inner-product score
$\alpha_{ij}$	Attention weights
$\mathbf{W}_s$	Static embeddings linear layer
$d_i^h$	Dynamic embeddings
$s_i^h$	Static embeddings
$o_i$	Output Hadamard layer
$\mathbf{W}_o, \mathbf{b}_o$	Output linear layer
$\mathcal{P}$	Scoring function
$\mathcal{K}(\cdot)$	Kernel function
$\mu$	Baseline intensity
$\alpha$	Weight parameter
$\mathcal{B}$	Negative sample size
$N$	Sampling for Monte-Carlo integration
$M$	Mini batch segment size

Table 5: Notations

$$\begin{aligned}
o_i &= \mathbf{W}_o^T (\mathbf{d}_i^h - \mathbf{s}_i^h)^2 + b_o, \\
o_{i'} &= \mathbf{W}_{o'}^T (\mathbf{d}_i^h - \mathbf{s}_{i'}^h)^2 + b_{o'}, \\
\mathcal{P}^h &= \frac{1}{k} \sum_{i=1}^k \log(1 + \exp(o_i)) + \frac{1}{k'} \sum_{i=1}^k \log(1 + \exp(o_{i'})).
\end{aligned} \tag{17}$$

Here,  $\mathbf{W}_o, \mathbf{W}_{o'} \in \mathbb{R}^{d \times 1}$  and  $b_{o'}, b_o \in \mathbb{R}$  are learnable parameters of output layer. In our model the value of  $f(\mathbf{v}_1(t), \mathbf{v}_2(t), \dots, \mathbf{v}_k(t)) = \mathcal{P}^h$ .

## D Datasets

**Email Network** [email-Enron (Klimt and Yang 2004); email-Eu (Paranjape, Benson, and Leskovec 2017)]. This is a sequence of timestamped email interactions between employees at a company. The entities involved are the email addresses of the sender and receivers. The timestamps in this are recorded at a resolution of 1-millisecond, which are scaled down in all our experiments. The scaling factor is chosen as the median value of interevent duration for both datasets.

**congress-bills** (Fowler 2006). Here, interactions are the legislative bills put forth in the House of Representatives and the Senate in the US. The entities involved in an interaction are the US congresspersons who sponsor and co-sponsor the bill. The timestamp in this is the date when the bill is introduced.

**Drug Networks [NDC-classes, NDC-sub]**. Each interaction is a sequence of drugs, and timestamps are the dates at which the drugs were introduced in the market. In NDC classes, nodes involved in the interaction are class labels applied to the drugs. In NDC-sub, nodes involved in the interaction are substances that make up the drug.

**[CastGenre, CastKeyword, CastCrew]** These datasets are prepared from the subset of Kaggle’s The Movie Dataset<sup>3</sup>. For the CastGenre dataset, nodes in the left hyperedge are movie actors and the genres associated with the movie in the right hyperedge. For the CastKeyword dataset, we use keywords associated with movie plots for right hyperedge instead of genres. Here, the movie release date is the timestamp associated with each hyperedge. For the CastCrew dataset, we use the name of the crew associated with movie production for the right hyperedge. We only considered movies released after 1990 and followed the same preprocessing as the recent work (Sharma et al. 2021) for preparing these datasets.

## E Baselines

**Rayleigh Hyperedge (RHE)**. In this, we keep embeddings  $\mathbf{v}(t)$  fixed for every time instance. Then intensity is modeled as the Rayleigh process as in Equation 4, and duration predictions are made using the closed form expression in Equation 9.

**Rayleigh Dynamic Hyperedge (RDHE)**. Similar to RHE, but here we allow the node embeddings  $\mathbf{v}(t)$  to evolve when an interaction involving the node  $v$  occurs as per the Interaction Update stage but not through Temporal Drift and History Aggregation stage. So,  $\mathbf{v}(t)$  is piece-wise continuous (node embeddings are constant during interevent time) as we do not allow the model to evolve the node embedding during the interevent time. This model is based on DeepCoevolve (Dai et al. 2016) model used for sequential recommendation.

**Dynamic Edge-Drift (DE-Drift)**. In this model, each hyperedge event is modeled as a concurrent pairwise edge events. For example,  $h = \{v_1, v_2, v_3\}$  is modeled as concurrent edge events  $\{(v_1, v_2), (v_2, v_3), (v_1, v_3)\}$ . Here, node embeddings  $\mathbf{v}(t)$  are allowed to evolve during the interevent time using Temporal Drift stage. For predicting interaction at time  $t$ , we use the product of all conditional intensity functions of concurrent edge events (e.g.  $\lambda_{v_1, v_2}(t), \lambda_{v_2, v_3}(t), \lambda_{v_1, v_3}(t)$  for interaction  $h$ ). For predicting the time of interaction, we calculate  $\hat{t}$  in Equation 8 for each edge in the interaction and average them. In this model, conditional intensity is defined as  $\lambda_{v_1(t), v_2(t)} = \mathbf{v}_1^T(t) \mathbf{v}_2(t)$ .

**Dynamic Edge (DE)**. This uses the same modeling technique as DE-Drift. In addition, it also uses Interaction Update stage to evolve embeddings.

**Dynamic Hyperedge-Drift (DHE-Drift)**. This uses hyperedge modeling explained in Section 2.2, but do not use Interaction Update and History Aggregation stage.

**Dynamic Hyperedge (DHE)**. This uses hyperedge modeling explained in Section 2.2, but does not use History Aggregation stage.

**Hypergraph Dynamic Hyperedge-hist (HGDHE-hist)**. This uses hyperedge modeling explained in Section 2.2, but does not use Interaction Update stage.

**Bipartite Dynamic Edge (BDE)**. This is the bipartite version of DE, where each bi-partite hyperedge is modeled as concurrent bipartite edges. For example,  $h = (\{v_1, v_2\}, \{v_1', v_2'\})$  is modeled as concurrent edge events  $\{(v_1, v_1'), (v_2, v_1'), (v_1, v_2'), (v_2, v_2')\}$ . For learning node representation, the left and right nodes have a different set of parameters as explained in Section 4.

**Bipartite Dynamic Hyperedge (BDHE)**. This is the bipartite version of DHE.

**DyReP**. (Trivedi et al. 2019) This is designed for forecasting pairwise links in a dynamic graph. The architecture of this is similar to DE model, however it uses graph neural network based neighborhood aggregation in the Interaction Update Stage and do not use Temporal Drift to evolve the embedding during the interevent time. We used our own implementation of this model to get comparable performance.

<sup>3</sup><https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata>

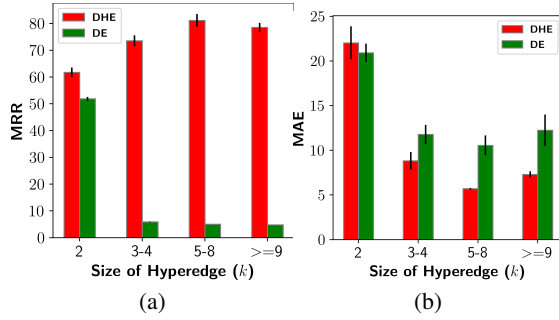


Figure 3: Figure 3a shows the effect of hyperedge size ( $k$ ) on interaction type prediction, and Figure 3b shows the effect on interaction duration prediction for email-Eu dataset

**TGAT.** (da Xu et al. 2020) This is a dynamic link prediction model for temporal graphs. For learning dynamic node embeddings, it uses graph attention to aggregate the topological features by considering the temporal history of the data. We transferred the hypergraph dataset to pairwise edge by clique expansion and used the code<sup>4</sup> provided by the authors to learn the model.

**TGN.** (Rossi et al. 2020) This model is for link prediction in temporal graphs. In this, interactions are stored in the form of messages and are used to update the respective nodes’ memory modules. This memory embedding is used as an input to the temporal graph attention module for getting the dynamic node embedding. We followed the same preprocessing as for TGAT and used the code<sup>5</sup> provided by the authors to learn the model.

## F Computational Infrastructure

We used a workstation with AMD EPYC 7742 processor, Nvidia-DGX-A100 with 512 GB of main memory, and 4 x A100 GPUs with 40 GB of memory.

## G Effect of Hyperedge Size on Performance

The results for analysis on hyperedge size on performance explained in Section 6.3 for datasets email-Eu, congress-bills, NDC-classes, and NDC-sub are shown in Figures 3, 4, 5, and 6, respectively. In all figures, we observe that MRR for interaction type prediction is poor for DE compared to DHE for hyperedge of size ( $k$ ) more than 2.

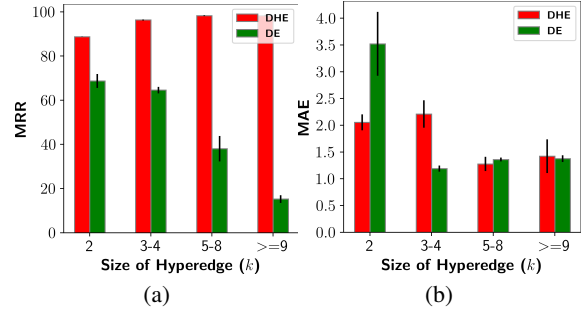


Figure 4: Figure 4a shows the effect of hyperedge size ( $k$ ) on interaction type prediction, and Figure 4b shows the effect on interaction duration prediction for congress-bills dataset

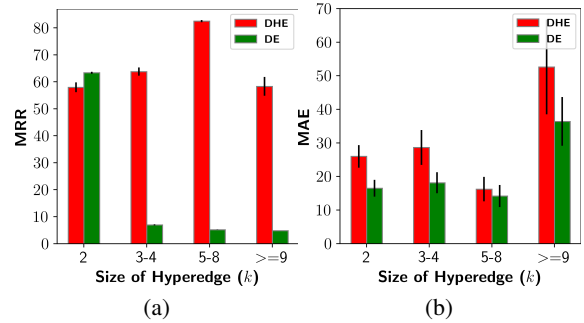


Figure 5: Figure 5a shows the effect of hyperedge size ( $k$ ) on interaction type prediction, and Figure 5b shows the effect on interaction duration prediction for Enron dataset

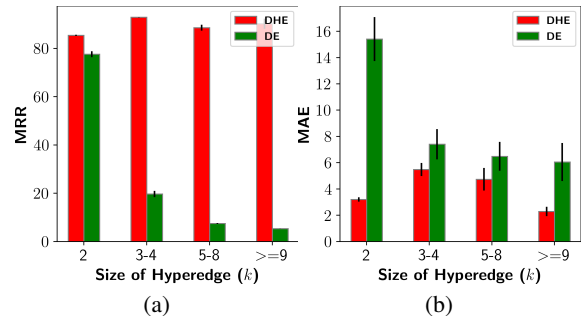


Figure 6: Figure 6a shows the effect of hyperedge size ( $k$ ) on interaction type prediction, and Figure 6b shows the effect on interaction duration prediction for NDC-sub dataset

<sup>4</sup><https://github.com/StatsDLMathsRecomSys/Inductive-representation-learning-on-temporal-graphs>

<sup>5</sup><https://github.com/twitter-research/tgn>