Enhancing Information Diffusion Prediction with Self-Supervised Disentangled User and Cascade Representations

Zhangtao Cheng*
Wenxue Ye*
Leyuan Liu†
zhangtao.cheng@outlook.com
wenxue.ye@outlook.com
leyuanliu@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Wenxin Tai†
Fan Zhou
wxtai@std.uestc.edu.cn
fan.zhou@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
Kash Institute of Electronics and Information Industry
Kashgar, Xinjiang, China

ABSTRACT

Accurately predicting information diffusion is critical for a vast range of applications. Existing methods generally consider user re-sharing behaviors to be driven by a single intent, and/or assume cascade temporal influence to be unchanged, which might not be consistent with real-world scenarios. To address these issues, we propose a self-supervised disentanglement framework (DisenIDP) for information diffusion prediction. First, we construct intent-aware hypergraphs to capture users' potential intents from different perspectives, and then perform the light hypergraph convolution to adaptively activate disentangled intents. Second, we extract long-term and short-term cascade influence via independent attention-based encoders. Finally, we set a self-supervised disentanglement task to alleviate the information loss and learn better-disentanglement representations. Extensive experiments conducted on two real-world social datasets demonstrate that DisenIDP outperforms state-of-the-art models across several settings.

CCS CONCEPTS

• **Information systems** → *Information systems applications.*

KEYWORDS

Information diffusion prediction, hypergraph representation, disentangled representation learning, self-supervised learning.

ACM Reference Format:
Zhangtao Cheng, Wenxue Ye, Leyuan Liu, Wenxin Tai, and Fan Zhou. 2023. Enhancing Information Diffusion Prediction with Self-Supervised Disentangled User and Cascade Representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615230

*Both authors contributed equally to this research.
†Corresponding author.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615230

1 INTRODUCTION

Information rapidly spreads to numerous users through posting and retweeting behaviors, resulting in a *cascade* of user activation. Predicting the *diffusion* of pieces of information can benefit analyzing how information diffuses among users [34], which has been playing an increasingly crucial role in numerous social applications, i.e., recommendation [2, 33], and popularity prediction [27, 30, 35].

Information Diffusion Prediction (IDP) can be principally summarized into three categories. (1) *Independent cascade models* [14, 18] try to oversimplify complexity of information propagation for IDP according to the independent diffusion assumption, which limit their applicability to real social scenarios. (2) *Feature engineering methods* [1, 7, 9, 32] aim to design and incorporate hand-crafted cascade features, which require abundant expert knowledge and thus are hardly generalized to new domains. (3) *Deep learning methods* [21, 25, 31] focus on combining sequence models (e.g., GRU [3] and attention layer [23]) and graph-based models (e.g., GCN [15] and hypergraph [8]) to design an automatic framework to capture the structural and temporal cascade representations for IDP.

Challenges: Although these methods have achieved promising performance, there are still two unresolved challenges: (1) *User intents behind re-sharing behaviors are entangled.* In the real-world diffusion process, user re-sharing behaviors are generally driven by multiple intents simultaneously, such as user interest and diffusion dependency. For example, a user retweets ChatGPT-related news probably because he/she is interested in ChatGPT or a fan of the news publisher. (2) *Cascade temporal influence is dynamic drifting.* In practice, cascade temporal influence [20] is not static, but dynamically drifts over time, containing stable Long-term influence and dynamic Short-term (LS-term) influence. For instance, the blogger of a distinctive topic has a fixed number of followers, which generally reveals the long-term influence. Meanwhile, information propagation tends to evolve rapidly and be impacted by hot events and recent user interests, which indicates the short-term influence.

Present work: We propose *DisenIDP*, a novel self-supervised **Disen**tanglement framework for **IDP**, to better disentangle user intents and LS-term temporal influence through self-supervision signals from cascades. First, we mine user intents underlying resharing behaviors and construct interest- and dependency-wise hypergraphs from user resharing history and cascades. DisenIDP encodes the potential high-order relations among users and obtain intent-specific user representations via multiple types of hypergraphs. To

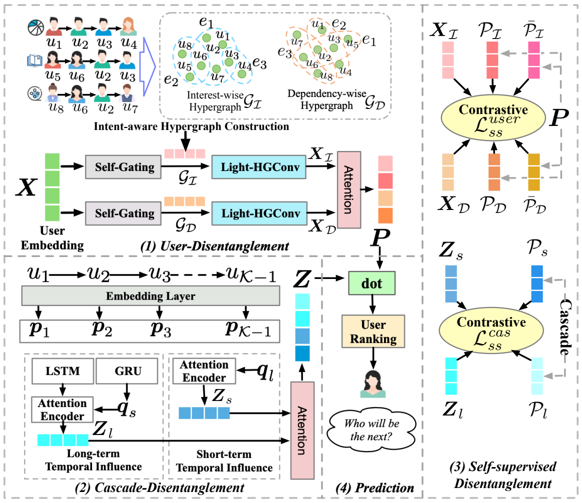

Figure 1: Overall framework of DisenIDP.

disentangle LS-term cascade temporal dependence, we design two separate attention-based encoders with different dynamics over time. To supervise the disentanglement process and prevent information loss, we design a self-supervised auxiliary task to guide disentanglement and learn fine-grained user and cascade representations. Finally, we integrate the IDP task and self-supervised disentanglement task within a primary & auxiliary learning framework, jointly optimizing them for improved diffusion prediction.

Contribution: Our contributions are threefold: **(1)** We consider coupled user intents and dynamic cascade temporal influence for IDP, and propose DisenIDP to take pioneer step of disentangling two kinds of representations at finer granularity. **(2)** We perform intent-aware hypergraph convolution on two intent-specific hypergraphs to generate disentangled intent embeddings, and then employ separately attention-based encoders to capture LS-term temporal influence. Finally, we construct a self-supervised disentanglement task to encourage disentanglement as well as alleviate the information loss issue. **(3)** Extensive experiments on two real cascade datasets show that our DisenIDP outperforms existing state-of-the-art baselines [1].

2 METHODOLOGY

Problem Definition: Let $\mathcal{U}$ and C denote the set of N users and M cascade items respectively. A cascade $c \in C$ is recorded as a time-order user activation sequence $c = \{(u_k^c, t_k^c) \mid u_k^c \in \mathcal{U}, t_k^c \in [0, \infty), k = 1 \ldots \mathcal{K}\}$, where $\mathcal{K}$ is the maximum cascade length. The tuple (u_k^c, t_k^c) means that user u_k^c reposted the current information at a certain timestamp t_k^c. Given an cascade c, the *goal* of IDP is to evaluate the activation likelihood $p(u_{\mathcal{K}+1}^c | c)$ for potential candidate user $u_{\mathcal{K}+1}^c \in \mathcal{U}$. An overview of our model is shown in Figure 1.

2.1 User-disentanglement

User Embedding. Existing methods principally parameterize each user in a holistic manner, failing to learn disentangled intent-specific representations [34]. Hence, we consider two-aspect user intents including *interest-aware* ($\mathcal{I}$), and *dependency-aware* ($\mathcal{D}$). Distinct from existing works, we design an *intent-aware self-gating* (ISG) operation to initialize the intent-specific user embeddings, which associates each gating layer with an intent. Formally, the ISG can be defined as: $\boldsymbol{X}_s = f_{gate}^s(\boldsymbol{X}) = \boldsymbol{X} \odot \text{sigmoid}(\boldsymbol{X}\boldsymbol{W}_s + \boldsymbol{b}_s)$, where $\boldsymbol{W}_s \in \mathbb{R}^{d \times d}$, $\boldsymbol{b}_s \in \mathbb{R}^d$ are intent-aware learnable parameters, $s \in \{\mathcal{I}, \mathcal{D}\}$ denotes different user intents, $\odot$ represents the element-wise product, $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ denotes the base user embedding matrix encoding users, and d is the adjustable latent dimensions. The self-gating mechanism designs a nonlinear gate [5] to generate user embeddings $\boldsymbol{X}^s$ indicating a certain user intent.

Intent-aware Hypergraph Construction. Generally, user resharing behaviors are triggered by the joint effect of shared interests and contextual relationships among previously activated users, which exists high-order relations among users. Furthermore, hypergraph [8] consists of hyperedges, which can connect an arbitrary number of nodes, and naturally describes complex high-order relations. Therefore, we construct two-aspect intent-aware hypergraphs $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, where $\mathcal{V}_s$ and $\mathcal{E}_s$ denote the set of nodes and hyperedges respectively, to independently describe the user high-order relations under certain intents.

• *Dependency-wise Hypergraph* $\mathcal{G}_{\mathcal{D}}$. The diffusion dependency [28], depicting who possibly infects whom, can partly reflect social relationships and propagation patterns of different users. To describe dependency relations among users, we first model each user as a hyperedge and then use w-size sliding window on cascades to observe local retweet processes and generate clear diffusion dependency relations according to all user sequences in the window.

• *Interest-wise Hypergraph* $\mathcal{G}_{\mathcal{I}}$. Users with shared interests generally induce correlated re-sharing behaviors without direct causal influence [17]. However, cascade sequences can't directly characterize user interests. To distill user interest-specific knowledge from cascades themselves, we model each cascade as a hyperedge in which users of retweeting the same cascade can be connected with each other. Different hyperedges, which are connected via shared users, integrally reveal user relations with shared interests.

Disentangled Hypergraph Convolution. Inspired by [26], we propose a *light hypergraph convolution* (LHGConv), which removes the activation function and feature transformation, to capture high-order information and obtain intent-specific user representations in the corresponding hypergraph. The LHGConv consists of a two-step aggregation process, i.e., *node-to-hyperedge* and *hyperedge-to-node*, for refining user representations. More formally, the propagation of LHGConv is defined as: $\boldsymbol{X}_s^{(l+1)} = \mathbf{D}_s^{-1}\boldsymbol{H}_s\mathbf{B}_s^{-1}\boldsymbol{H}_s^{\mathrm{T}}\boldsymbol{X}_s^{(l)}$, where $\boldsymbol{X}_s^{(0)}$ is initialized with embeddings $\boldsymbol{X}_s$, $\mathbf{D}_s$ and $\mathbf{B}_s$ denote node and hyperedge degree matrices, respectively. The hypergraph $\mathcal{G}_s$ can be described by an incidence matrix $\boldsymbol{H}_s$. After the hypergraph convolution operation, we obtain the user representations $\boldsymbol{X}_s^{(l+1)}$ integrating fine-grained certain intent semantic information from each hypergraph. Finally, we combine the user embeddings of L layers to avoid the over-smoothing [10]: $\bar{\mathrm{X}}_s = \frac{1}{L+1}\sum_{l=0}^{L} \mathrm{X}_s^l$.

Multi-intent Aggregation. Since different user intents show different importance for final diffusion prediction, we develop an attention mechanism [23] to selectively integrate the important information from two intent-specific user embeddings and generate comprehensive user embeddings representing distinct user intents.

[1] https://github.com/CZ-TAO12/DisenIDP

Formally, the attention layer is presented as follows:

$$\alpha_s = \frac{\exp\left(\boldsymbol{a}^\top \cdot \boldsymbol{W}_a \bar{\mathbf{X}}_s\right)}{\sum_{s' \in \{\mathcal{D},\mathcal{I}\}} \exp\left(\boldsymbol{a}^\top \cdot \boldsymbol{W}_a \bar{\mathbf{X}}_{s'}\right)}, \tag{1}$$

where $\boldsymbol{a} \in \mathbb{R}^d$ and $\boldsymbol{W}_a \in \mathbb{R}^{d\times d}$ are trainable parameters. Finally, the comprehensive intent-specific user representations can be computed as $\boldsymbol{P} = \sum_{s\in\{\mathcal{D},\mathcal{I}\}} \alpha_s \bar{\mathbf{X}}_s$.

2.2 Cascade-disentanglement

Diffusion prediction relies heavily on temporal influence [20], which describes the dynamic changes in user transitions and cascade content with the evolution of diffusion. To capture expressive cascade representations, we propose to disentangle cascades' long and short-term temporal influence. Since multi-aspect user intents trigger user re-sharing behaviors and affect cascade evolution processes, we use the intent-specific user feature $\boldsymbol{P}$ to initialize the activated users joining each cascade and consider the diffusion patterns of intent driving information propagation.

Long-term Temporal Influence Encoder (LTIE). The source user, *the publisher of information*, implies the topic and diffusion pattern of cascades [28], which can partly reflect long-term influence. To associate each user in the cascade with a corresponding root, we propose a source-aware attention encoder to learn contextual long-term influence representations. We first use the source u_1 embedding $\boldsymbol{p}_{u_1}$ as *a query* and then employing previously activated users as *keys* of the attention encoder. Formally, given a cascade c, the attention score of user u_j can be computed based on the source user u_1 as follows:

$$\alpha_j = \frac{\exp(\langle \boldsymbol{W}_l^q \boldsymbol{p}_{u_1}, \boldsymbol{W}_l^k \boldsymbol{p}_{u_j}\rangle)}{\sum_{j=1}^{\mathcal{K}-1} \exp(\langle \boldsymbol{W}_l^q \boldsymbol{p}_{u_1}, \boldsymbol{W}_l^k \boldsymbol{p}_{u_j}\rangle)}. \tag{2}$$

As a result, we obtain the learned cascade long-term representations $z_l^{\mathcal{K}}$ that is computed via a weighted aggregation of the entire cascade sequence: $z_l^{\mathcal{K}} = \sum_{j=1}^{\mathcal{K}-1} \alpha_j \boldsymbol{W}_l^v \boldsymbol{p}_{u_j}$. Furthermore, $\boldsymbol{W}_l^q$, $\boldsymbol{W}_l^k$ and $\boldsymbol{W}_l^v$ denote transformation matrices.

Short-term Temporal Influence Encoder (STIE). Since the sequential model can explore the trend of dynamic evolution to reveal the users' present sharing motivation, we consider sequential patterns of user transitions in short-term temporal influence modeling and design a sequential-aware attention encoder on top of a sequential model. Specially, we first feed the activated user sequence to Gate Recurrent Unit (GRU) [3] and then use the outputs as *query*. Furthermore, we employ another recurrent neural network (RNN) to model user sequences and then regard the outputs of RNN as the *keys*. Analogously, according to Eq. 2 and the weighted aggregation operation, we obtain learned cascade short-term features $z_s^{\mathcal{K}}$.

Multi-Head-Enhanced cascade Representation. Following Eq. (1), we obtain the comprehensive cascade representations $\boldsymbol{Z}^*$ containing user intent-specific and LS-term temporal information. To enhance cascade encoders with the capability of jointly attending multi-dimensional dependencies among users, we design a multi-head cascade encoder projecting the $\boldsymbol{Z}^*$ into Q latent feature spaces and performing head-specific attentive operations in parallel.

$$\begin{aligned} z_{k,q} &= \mathrm{ATTENTION}(z_k^* \boldsymbol{W}_q^{\text{query}}, z_k^* \boldsymbol{W}_q^{\text{key}}, z_k^* \boldsymbol{W}_q^{\text{value}}), \\ z_k &= (z_{k,1} || z_{k,2} || \dots || z_{k,q} || \dots || z_{k,Q}) \boldsymbol{W}^O, \end{aligned} \tag{3}$$

where $\mathbf{W}_q^{\text{query}}, \mathbf{W}_q^{\text{key}}, \mathbf{W}_q^{\text{value}} \in \mathbb{R}^{d/Q\times d}$, $\mathbf{W}^O \in \mathbb{R}^{d\times d/Q}$ are learnable parameter matrices, $z_k^* \in Z^* = \{z_k^* | k = 1 \dots \mathcal{K}\}$, and Q is the number of attention heads.

2.3 Self-supervised Disentanglement

In the disentanglement process, it is hard to obtain labeled data to explicitly supervise the user and cascade disentanglement. Meanwhile, due to the aggregation operations (e.g., Eq. 1), it results in a loss of fine-grained representation and leads to model sub-optimization. To fully utilize disentangled information and enhance the disentanglement process, we design an auxiliary task, a hierarchical self-supervised disentanglement learning [16, 22] (i.e., *user-wise discrimination* and *cascade-wise discrimination*), to enhance diffusion prediction (primary task). Specially, we design *proxies* serving as labels for user intents and cascade LT-term influence, and employ contrastive learning between the encoder outputs and proxies for better disentanglement in a self-supervised way.

• *User-wise Discrimination* (UD). We design two readout functions to generate *proxies* ($\mathcal{P}$) for user discrimination. Specifically, one readout function considers the node-level feature of each hypergraph, which is computed as $\boldsymbol{p}^s_{\mathcal{P},u} = \frac{\boldsymbol{P}^s_* \boldsymbol{H}^u_s}{\mathrm{sum}(\boldsymbol{H}^s_u)}$, where $\boldsymbol{P}^s_* = f^s_{gate}(\boldsymbol{P})$ avoids overfitting and mitigates gradient conflict between the primary and auxiliary tasks. $\boldsymbol{H}^u_s$ is the row vector of $\boldsymbol{H}_s$ according to the user u. Another readout function employs average pooling to summarize the obtained node-level hypergraph features into a graph-level feature $\bar{\boldsymbol{P}}^s_{\mathcal{P}} = \mathrm{AveragePooling}(\boldsymbol{P}^s_{\mathcal{P}})$. Finally, we use the pairwise ranking loss [13] as our learning objective to maximize the hierarchical mutual information:

$$\mathcal{L}^{user}_{ss} = -\sum_{s\in\{\mathcal{D},\mathcal{I}\}} \log \sigma\left(f_D\left(\boldsymbol{X}_s, \mathcal{P}_s\right) - f_D(\boldsymbol{X}_s, \tilde{\mathcal{P}}_s)\right), \tag{4}$$

where $f_D(\cdot) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is the discriminator function that is implemented as the dot product between two representations. Furthermore, $\tilde{\mathcal{P}}_s$ (or $\tilde{\boldsymbol{X}}_s$) is the negative sample obtained by corrupting $\mathcal{P}_s$ (or $\boldsymbol{X}_s$) with row-wise and column-wise shuffling. Note that the $\mathcal{P}_s$ can be replaced with any proxies, i.e., $\boldsymbol{P}^s_{\mathcal{P}}$ and $\bar{\boldsymbol{P}}^s_{\mathcal{P}}$.

• *Cascade-wise Discrimination* (CD). We design two methods to generate *proxies* for LS-term temporal influence. Since future-activated users better reflect long-term cascade evolution, we use source-aware attention in LTIE to generate a proxy $Z_{\mathcal{P},l}$ for the long-term influence. In addition, we use the convolution operation, where the size of the convolution kernel is w_2, to generate a proxy $Z_{\mathcal{P},s}$ for short-term representation. Following Eq. 4, we obtain the learning objective function $\mathcal{L}^{cas}_{ss}$ for cascade LS-term influence.

2.4 Diffusion Prediction

For the final prediction of the information diffusion, the probabilities $\hat{\mathbf{y}}_k \in \mathbb{R}^{N\times 1}$ for all users are calculated by: $\hat{\mathbf{y}}_k = \mathrm{softmax}(\boldsymbol{Z}^* \boldsymbol{P}^\top + \mathbf{M}_{\text{mask}})$, where $\mathbf{M}_{\text{mask}}$ is used to mask users who have already been activated. We adopt the cross entropy loss $\mathcal{L}_{\text{cross}}$ as the objective to optimize the primary task. Finally, we unify the objectives of

Table 1: Statistics of the datasets.

Dataset	#Users	#Cascades	#Train	#Val	#Test	Avg.Length
Twitter	12,627	3,454	2,763	345	346	38.22
Weibo	26,537	35,070	28,056	3,507	3,507	29.41

IDP (primary) and the task of maximizing hierarchical mutual information (auxiliary) for joint learning. The overall objective is defined as: $\mathcal{L} = \mathcal{L}_{\text{cross}} + \beta\mathcal{L}_{ss}^{user} + \lambda\mathcal{L}_{ss}^{cas}$, where β and λ are a hyper-parameter used to control the effect of the auxiliary task.

3 EXPERIMENTS

3.1 Experimental Settings

Datasets. We conduct experiments to evaluate DisenIDP on **Weibo** [4] and **Twitter** [11]. Detailed statistics are shown in Table 1. Following existing works [21, 29], we take each information content (i.e., URL and blog) and its retweeting users as an independent diffusion process (i.e., an cascade).

Metrics. Following [29, 31], we employ two widely used ranking metrics: MAP@K (M@K) and Hits@K (H@K), K = [10, 100].

Baselines. We compare our model with the following eight strong baselines: (1) Sequential-based methods: DeepDiffuse [12] and NDM [28]. (2) Graph-based methods: Topo-LSTM [24], SNIDSA [25], FOREST [29], Inf-VAE [19], DyHGCN [31], and MS-HGAT [21].

Parameter Settings. Our proposed DisenIDP and all baselines are tuned to the best performance with early stopping. For experimental results, we run each model on each dataset five times and report the mean performance. The dimension of the hidden unit is set to 64 for all baselines. Other parameters of baselines follow the recommended settings in original papers. The batch size is set to 64 in the training process. The number of LSTM units is 200. The sliding window's size w is set to 10. The attention head Q is 4.

3.2 Evaluation Results

Overall Performance. Table 2 reports overall comparison results. We have the following observations: **(O1)** We can observe that DisenIDP significantly outperforms eight baselines consistently on two datasets. Specifically, on Weibo dataset, the performance of DisenIDP improves the best baseline by 14.23% and 10.66% in terms of H@10, and M@100, respectively. These results verify the effectiveness of our designs that disentangling user intents and cascade LS-term influence is critical for IDP. **(O2)** Sequential-based methods perform markedly worse than their counterparts. It is because they simply learn short-term user correlations following the sequential assumption and ignore the complex latent factors behind user re-sharing behaviors, such as user intents and cascade LS-term influence. **(O3)** Graph-based models, which additionally consider social topology among users, tend to outperform sequential-based models, not for DisenIDP. It shows that they only model coarse-grained Spatio-temporal cascade knowledge and can not distinguish complex relations between user and cascade content. Therefore, our design for disentangling user and LS-term temporal representations from cascades themselves is necessary.

Ablation Study. To verify the effectiveness of the key compositions in DisenIDP, we conduct ablation studies shown in Table 3.

(1) Multi-type Hypergraph. We compare the performance without dependency-wise hypergraph (w/o $\mathcal{G}_{\mathcal{D}}$), and interest-wise hypergraph (w/o $\mathcal{G}_{\mathcal{I}}$), respectively. We find that removing any type of

Table 2: Performance comparisons on two datasets.

Model	Twitter				Weibo			
	H@10	H@100	M@10	M@100	H@10	H@100	M@10	M@100
DeepDiffuse	5.72	21.61	5.93	6.99	0.74	5.73	0.23	0.36
Topo-LSTM	10.45	25.42	9.51	14.68	1.86	12.89	0.60	0.90
NDM	22.45	35.12	15.59	16.03	9.85	39.31	4.05	4.95
SNIDSA	25.67	43.59	16.34	18.89	10.73	39.51	4.75	5.52
FOREST	30.28	50.12	21.45	22.36	15.59	52.55	7.55	8.63
Inf-VAE	14.93	46.42	19.83	21.82	10.37	38.05	5.90	6.58
DyHGCN	32.78	58.53	21.57	22.45	14.65	51.65	7.13	8.27
MS-HGAT	29.12	56.68	16.44	17.37	12.67	40.05	6.50	7.38
DisenIDP	**34.01**	**60.39**	**23.04**	**23.94**	**17.81**	**57.40**	**8.23**	**9.55**
% Improv.	3.75	3.17	6.81	6.63	14.23	9.22	9.00	10.66

Table 3: Ablation study of DisenIDP.

Model		Twitter		Weibo	
		H@100	M@100	H@100	M@100
DisenIDP	**All**	**60.39**	**23.94**	**57.40**	**9.55**
Multi-type Hypergraph	w/o $\mathcal{G}_{\mathcal{D}}$	59.52	23.26	56.43	8.89
	w/o $\mathcal{G}_{\mathcal{I}}$	59.19	23.04	56.37	8.77
User-Disentanglement	HyperGAT	58.93	22.98	55.55	8.93
	HyperGCN	58.89	22.88	54.21	8.76
Cascade-Disentanglement	w/o LTIE	55.20	17.62	44.98	7.34
	w/o STIE	57.59	19.95	45.35	7.44
Self-supervised-Disentanglement	w/o UD	59.63	23.35	56.63	8.81
	w/o CD	60.02	23.75	56.42	8.79

hypergraph would not lead to severe performance degradation. This indicates that capturing high-order relations between users and cascades via hypergraph is helpful for enhancing the model's generalization. Moreover, the combination of dependency-wise and interest-wise hypergraphs leads to further improvement, which validates the effectiveness of disentangling users' multiple intents.

(2) User-Disentanglement. When we use different propagation mechanisms for hypergraph structure learning, i.e., HyperGAT [6], HyperGCN [8], the performance results show that the LHGConv is suitable for IDP and achieves excellent performance.

(3) Cascade-Disentanglement. When we remove LTIE or STIE modules, the performance drops significantly. This result verifies the fact that the cascade LS-term temporal influence exists in the diffusion process and is essential for cascade modeling.

(4) Self-supervised Disentanglement. We build a variant without UD (w/o UD) or CD (w/o CD) to investigate the efficacy of auxiliary tasks. We observe that DisenIDP with auxiliary tasks outperforms other variants, highlighting the usefulness of auxiliary tasks in facilitating disentanglement.

4 CONCLUSION

We introduced DisenIDP, a self-supervised disentanglement framework to separate user intents and cascade LS-term temporal influence. DisenIDP was evaluated on two real-world cascade datasets and outperformed eight competitive baselines, indicating the benefits of disentanglement for information diffusion prediction. Further research is needed to explore a generalized disentanglement framework for different information cascade prediction tasks.

5 ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (Grant No.62072077 and No.62176043), and Natural Science Foundation of Sichuan Province (Grant No.2022NSFSC0505).

REFERENCES

[1] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2013. Influence-based network-oblivious community detection. In *ICDM*. 955–960.

[2] Zhangtao Cheng, Joojo Walker, Ting Zhong, and Fan Zhou. 2022. Modeling multi-view interactions with contrastive graph learning for collaborative filtering. In *IJCNN*. 1–8.

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[4] PEOPLE'S DAILY. 2022. Weibo Dataset. http://data.sklccc.com/2022/.

[5] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *ICML*. 933–941.

[6] Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. 2020. Be More with Less: Hypergraph Attention Networks for Inductive Text Classification. In *EMNLP*. 4927–4936.

[7] Nan Du, Le Song, Manuel Gomez Rodriguez, and Hongyuan Zha. 2013. Scalable influence estimation in continuous-time diffusion networks. *Nips* 26 (2013).

[8] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph Neural Networks. In *AAAI*. 3558–3565.

[9] Manuel Gomez-Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the Temporal Dynamics of Diffusion Networks. In *ICML*, Lise Getoor and Tobias Scheffer (Eds.). 561–568.

[10] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*. 639–648.

[11] Nathan O Hodas and Kristina Lerman. 2014. The simple rules of social contagion. *Scientific reports* 4, 1 (2014), 1–7.

[12] Mohammad Raihanul Islam, Sathappan Muthiah, Bijaya Adhikari, B Aditya Prakash, and Naren Ramakrishnan. 2018. Deepdiffuse: Predicting the'who'and'when'in cascades. In *ICDM*. 1055–1060.

[13] Mete Kemertas, Leila Pishdad, Konstantinos G Derpanis, and Afsaneh Fazly. 2020. Rankmi: A mutual information maximizing ranking loss. In *CVPR*. 14362–14371.

[14] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *SIGKDD*. 137–146.

[15] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

[16] Haoyang Li, Xin Wang, Ziwei Zhang, Zehuan Yuan, Hang Li, and Wenwu Zhu. 2021. Disentangled contrastive learning on graphs. *Neurips* 34 (2021), 21872–21884.

[17] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[18] Manuel Gomez Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. 2014. Uncovering the structure and temporal dynamics of information propagation. *Network Science* 2, 1 (2014), 26–65.

[19] Aravind Sankar, Xinyang Zhang, Adit Krishnan, and Jiawei Han. 2020. Inf-vae: A variational autoencoder framework to integrate homophily and influence in diffusion prediction. In *WSDM*. 510–518.

[20] Cosma Rohilla Shalizi and Andrew C Thomas. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research* 40, 2 (2011), 211–239.

[21] Ling Sun, Yuan Rao, Xiangbo Zhang, Yuqian Lan, and Shuanghe Yu. 2022. MS-HGAT: Memory-enhanced Sequential Hypergraph Attention Network for Information Diffusion Prediction. In *AAAI*.

[22] Wenxin Tai, Tian Lan, Zufeng Wu, Pengyu Wang, Yixiang Wang, and Fan Zhou. 2023. Improving session-based recommendation with contrastive learning. *UMUAI* 33, 1 (2023), 1–42.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[24] Jia Wang, Vincent W Zheng, Zemin Liu, and Kevin Chen-Chuan Chang. 2017. Topological recurrent neural network for diffusion prediction. In *ICDM*. 475–484.

[25] Zhitao Wang, Chengyao Chen, and Wenjie Li. 2018. A sequential neural information diffusion model with structure attention. In *CIKM*. 1795–1798.

[26] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *ICML*. 6861–6871.

[27] Xovee Xu, Fan Zhou, Kunpeng Zhang, Siyuan Liu, and Goce Trajcevski. 2021. Casflow: Exploring hierarchical structures and propagation uncertainty for cascade prediction. *IEEE TKDE* (2021).

[28] Cheng Yang, Maosong Sun, Haoran Liu, Shiyi Han, Zhiyuan Liu, and Huanbo Luan. 2019. Neural diffusion model for microscopic cascade study. *IEEE TKDE* 33, 3 (2019), 1128–1139.

[29] Cheng Yang, Jian Tang, Maosong Sun, Ganqu Cui, and Zhiyuan Liu. 2019. Multi-scale Information Diffusion Prediction with Reinforced Recurrent Networks. In *IJCAI*. 4033–4039.

[30] Liu Yu, Xovee Xu, Goce Trajcevski, and Fan Zhou. 2022. Transformer-enhanced Hawkes process with decoupling training for information cascade prediction. *Knowledge-Based Systems* 255 (2022), 109740.

[31] Chunyuan Yuan, Jiacheng Li, Wei Zhou, Yijun Lu, Xiaodan Zhang, and Songlin Hu. 2020. DyHGCN: A dynamic heterogeneous graph convolutional network to learn users' dynamic preferences for information diffusion prediction. In *ECML-PKDD*. 347–363.

[32] Jing Zhang, Jie Tang, Yuanyi Zhong, Yuchen Mo, Juanzi Li, Guojie Song, Wendy Hall, and Jimeng Sun. 2017. Structinf: Mining structural influence from social streams. In *AAAI*.

[33] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52, 1 (2019), 1–38.

[34] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.

[35] Fan Zhou, Liu Yu, Xovee Xu, and Goce Trajcevski. 2021. Decoupling representation and regressor for long-tailed information cascade prediction. In *SIGIR*.