A Extended Related Work

A.1 Citation/Cascade prediction

Citation prediction is a vital sub-task of automatic academic evaluation as it enables the estimation of potential impact. It is closely examined within cascade prediction tasks, which share analogous graph structures and objectives [60]. The primary goal of cascade prediction is to anticipate a post's popularity based on user interactions. Similarly, in the context of citation networks, the paper can serve as the post, while the authors can serve as the users. Citation count prediction focuses on individual papers' content and distinctive features, heavily relying on comprehensive feature extraction techniques. In contrast, cascade prediction delves into the intricate relationships between various entities, including papers and authors, directly exploiting graph structure information.

Their approaches can be broadly classified into three categories: stochastic models, feature-based models, and deep learning models. Stochastic models can predict future citation counts by fitting the curve of citation counts [14], following Zipf-Mandelbrot's law [38]. Recently, machine learning models have shown promising results by utilizing manually extracted features from various meta-data (*e.g.* content, author, venue) [11, 16, 34, 41, 51, 55]. More presently, deep neural networks have dominated this task by applying advanced Natural Language Processing (NLP) and Computer Vision (CV) methods to extract abstractive representations from paper content [1, 9, 19, 49]. For cascade prediction, graph embeddings, sequence models, and GNN models are employed to extract structural information from the underlying graph, and then temporal information is encoded with sequence models [5, 6, 8, 25, 48, 52].

However, most existing models for citation and cascade prediction exhibit sub-optimal performance. They fail to fully exploit the valuable information presented in paper content and scientific context within the citation network simultaneously. Furthermore, their strict data selection and random data splitting strategies may hinder their practicality. Many valuable samples will be filtered, and the temporal information within the dynamic context will be neglected. Thus, models in this setting are more prone to face distribution shift issues. In contrast, our study revises the problem setting of impact prediction in the dynamic context [17, 50]. We ensure non-overlapping training and testing observation windows, retaining the real data distribution without sample filtering. We also utilize both content and dynamic context, introducing disentangled representation learning to provide a more practical impact estimation.

A.2 Dynamic/Heterogeneous Graph Neural Networks

GNNs [15, 36, 39, 47] are widely applied to handle non-Euclidean data like graphs. Recently, research interest has surged in dynamic graphs, which incorporate temporal information, as well as in heterogeneous graphs, which involve multiple node types or edge types. These types of graphs are more prevalent in practical scenarios and encapsulate more intricate information compared to static homogeneous graphs, driving the formulation and refinement of diverse methodologies within the academic realm.

The dynamic graph can be divided into multiple snapshots at different time points. Previous models like DGCN [31], Dysat [35], and ROLAND [54] typically encode the snapshots with static GCN as the structural encoder, followed by sequential models like Recurrent Neural Networks (RNN) or Transformer Encoder. Moreover, certain models incorporate GCN into RNN models by configuring the parameters within RNN cells as GCN parameters [33]. Within heterogeneous graphs, on one side, R-GCN serves as a prominent model with its separate message-passing architecture [37]. It conducts independent message passing within different relations and aggregates them to update node representations. On the other side, the relation-learning method transforms heterogeneous graphs into homogeneous ones, while retaining node and edge type indicators to preserve heterogeneous information [18, 54]. An example of this technique is HGT [18], which utilizes multi-head attention modules with specific parameters tailored to different node and edge types, effectively modeling the complexity of heterogeneous graphs.

Citation networks can be regarded as dynamic and heterogeneous graphs as they evolve over time and encompass diverse entities such as papers, authors, and venues. Addressing the complexities within this network involves two primary challenges: (1) How to encode the temporal target-centric information within a single snapshot? (2) How to model distinct characteristics of information diffusion in citation networks? To overcome these challenges, our study proposes a Citation-aware GNN Encoder that effectively models both dynamic and heterogeneous graphs of target papers. These graphs are constructed from the citation network annually and incorporate multiple metadata nodes beyond the papers themselves. (1) To encode the temporal target-centric information within a single snapshot, we employ a type-aware attention readout. It aggregates snapshot information considering the paper type and publication time. (2) To model distinct characteristics of information diffusion in citation networks, we propose a novel GCN module CompGAT. This module focuses on capturing comparisons and co-cited/citing information between papers, drawing inspiration from bibliometrics theories [46].

A.3 Disentangled Representation Learning

Disentangled representation learning, aiming to discern and separate fundamental explanatory factors [3], stands as a pivotal method within deep learning. It strives to produce resilient, manageable, and interpretable representations. Existing research predominantly focuses on the field of CV, often employing methods like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to disentangle relevant latent factors [4, 10, 24, 28, 30]. Researchers explore techniques to align distinct representations with specific factors through explicit and implicit guidance, employing approaches like direct labeling, loss penalties, and regularization. For instance, DR-GAN [42] utilizes an adversarial network to generate diverse facial poses, aiming to disentangle pose information from facial attributes. In NLP, disentangled representation learning finds applications in various tasks like generation. For example, John et al. [21] leverage VAE to disentangle text

semantics from its writing style, enabling independent manipulation of text semantics and style. This separation allows precise control over content and style in text generation.

Recently, there has been a growing interest in applying disentangled representation learning to graph learning tasks. Given that graph data encompasses both structural and attributed information, numerous new methodologies have emerged to address this complexity. In the feature space, approaches inspired by capsule networks have emerged to partition node features into multiple hidden channels for implicit disentangled modeling [29]. Some methods introduce additional constraints to enable effective separation [27]. Moving into the structural space, subsequent models strive to factorize the input graphs into multiple subgraphs to facilitate distinct message-passing strategies [26, 53]. Recent models have expanded their scope to isolate causal and biased information presented in both the graph structure and node features, with a particular emphasis on causal effects [12, 40]. These models commonly employ techniques to generate masks within the adjacency matrix and node features. This process aims to identify the crucial elements of the graph within the structural and attributed spaces. In addition, certain studies extend disentangled GNNs to specialized scenarios such as recommendation systems, handling intricate graph structures in heterogeneous or dynamic graphs [44, 45, 56–58]. Besides the masking strategy considering solely the graph structures and node features, they integrate classical methods like VAE and adversarial learning from CV and NLP to handle task-specific information from a broader perspective.

In contrast to previous studies, our approach acknowledges that the potential impact within the citation network is influenced not only by the paper's contribution but also by factors related to its popularity. These popularity factors encompass the amplifying effect owing to the highly-cited paper nodes in information diffusion and the manifestation of the Matthew effect through collective conformity. Following previous disentangled studies in recommendations [7, 59], we consider the final predicted values as mixed results of multiple factors and introduce auxiliary tasks for separate encoding of different perspectives. Based on the dynamic heterogeneous graph, we disentangle the potential impact into diffusion, conformity, and contribution values, seeking a more reasonable estimation in explicit modeling. Specifically, we devise dedicated auxiliary tasks to extract values associated with these respective factors.

B Notation Table

Table 1: The complete notation table of our paper.

Notation	Description
N	All predicted samples
G	Input dynamic heterogeneous graph
t	The t-th time point of the whole time window
G^t	The t-th subgraph of the input graph
ϵ^t, v^t	Set of nodes and edges at t-th time point
ϕ, φ	Type map function of nodes and edges
U, R	Set of node types and edge types
p	The paper node type
$\boldsymbol{h}_{ti,u}^{(l)}$	The i-th node hidden representation of type u at t-th time step in l-th layer
$\boldsymbol{H}_s$	Snapshot representation matrix containing all $h_{ti,s}$ of the i-th sample
AGG	Aggregation functions of update representations from multiple relationships (edges)
f_r	Message passing method (module) of the edge r
g_r	Subgraph of the edge r
$\boldsymbol{h}_{r_{src}}, \boldsymbol{h}_{r_{dst}}$	The source and target node representations of the edge r
$\boldsymbol{W}$	The model parameters
$\boldsymbol{C}$	Normalized co-cited/citing strength matrix
$\boldsymbol{o}_i$	The final encoded representation of the i-th target sample before disentanglement
$\boldsymbol{o}_i^{dif}$, $\boldsymbol{o}_i^{con}$	The representation of the i-th sample's diffusion and conformity view
$\boldsymbol{z}_i^{ori \vee pos \vee neg}$	The i-th sample's projected representation of the original, positive and negative view
$\mathcal{L}_{dif}$	Objective to disentangle value of diffusion view
$\mathcal{L}_{con}$	Objective to disentangle value of conformity view
$\mathcal{L}_{ort}$	Orthogonal constraint to ensure the non-overlapping information of disentangled views
$\mathcal{L}_{reg}$	Objective to predict the citation increments as the main task
L	The number of stacked graph encoder
λ	The importance of co-cited/citing strengths
β	The weight of the whole disentanglement task

C Complexity Analysis

The primary complexity lies in the graph encoder component of our model, CGE. Within each layer of CGE, multiple R-GCNs operate on snapshots captured at different time points, alongside a snapshot representation updater and a multi-layer transformer encoder. For edges apart from "cites" and "is cited by" in R-GCNs, a simple MLP ($W^u \in \mathbb{R}^{d\times d}$) serves as learnable parameters for GIN, with a complexity of $O\left(LT(|R|-2)(|v^t_{r_{src}}|d^2+|\epsilon^t_r|d)\right)$. For "cites" and "is cited by" edges, the backbone of CompGAT, GATv2, incurs a complexity of $O\left(LT(|v^t_p|d^2+|\epsilon^t_p|d)\right)$. Due to the modeling of concatenated source and target representations, it further requires $O\left(LT|\epsilon^t_p|d^2)\right)$ time, resulting in an overall complexity of $O\left(LT(|v^t_p|+|\epsilon^t_p|)d^2\right)$. The snapshot representation updater, functioning as a cross-attention module, operates with a complexity of $O\left(LT(|v^t_p|d+d^2)\right)$, while the temporal encoder exhibits a complexity of $O\left(L(T^2d+Td^2)\right)$. These modules can be readily disregarded compared to R-GCNs. For PDE, which solely models the final target representation using multiple MLPs ($O\left(d^2\right)$), its contribution can also be omitted. Thus, the resulting complexity of DPPDCC is $O\left(LT(|v^t|d^2+|\epsilon^t_p|d^2+|\epsilon^t|d)\right)$. The final time complexity depends on the sampling strategy for graph size and the sparsity of *paper* node subgraphs. Since citation graphs are sparse and only a small portion of the entire graphs is sampled, computations are confined mainly within the limited *paper* subgraphs. Therefore, DPPDCC is comparable to other dynamic heterogeneous GNNs.

D Experimental Details

D.1 Datasets

Table 2: Detailed statistics for ground-truth samples, nodes, and edges of the citation network, as well as sample information of training, validation, and testing sets.

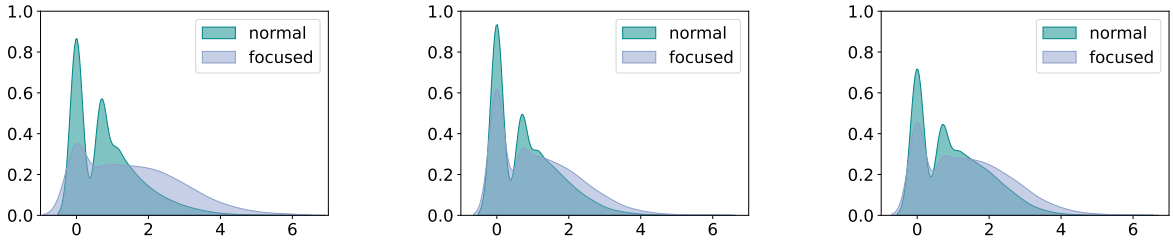
dataset		CS	CHM	PSY
ground-truth sample		2,513,197	1,818,138	1,857,277
node	paper	1,628,853	1,376,599	1,297,771
	author	1,598,925	1,946,073	1,585,595
	venue	12,524	8,389	13,775
	time	147	187	202
edge	cite	11,534,431	10,382,698	13,401,112
	write	5,123,460	6,355,630	4,813,135
	publish	1,566,442	1,318,158	1,230,440
	have	1,628,853	1,376,599	1,297,771
sample	train (2012)	183,105	227,304	219,499
	val (2014)	224,086	256,403	250,229
	test (2017)	300,000	300,000	300,000
	qualified pool	1,112,611	1,093,660	1,015,579

Our dataset is constructed using S2AG [23], a comprehensive repository housing roughly 100 million scientific publications, offering extensive metadata and unique identifiers. To enhance experimental efficiency and accommodate the diverse characteristics across domains, we organize these publications into subsets based on their respective fields of study. To establish the global citation network, we locate English papers with complete metadata (title, abstract, author IDs, and venue ID) or those of high impact despite lacking venue information. From the dataset, considering the data quality and the disciplinary differences, we select three fields: computer science (CS), chemistry (CHM), and psychology (PSY). Their total number of papers and the completeness of their metadata are the highest among all fields. Additionally, the differences in citation patterns resulting from varying research paradigms, such as variations in citation half-lives [32], render these fields appropriate representative samples for demonstrating citation differences across disciplines. Relevant statistics are summarized in Table 2.

To select samples for testing, we target papers published before the test observation point. These papers should be written in English, have complete metadata, and include at least one reference without imposing constraints on the number of citations received to maintain the integrity of the data distribution. From the eligible pool, we randomly sampled 300,000 papers to compose the test set. Additionally, we gather all the associated metadata for these papers, encompassing titles, abstracts, authors, venues, and publication times. To imitate the practical scenario, the training observation point is set 5 years before the test observation point (equal to the observation time window T and predicted interval Δ) for non-overlapping. The validation observation point precedes the test observation point by 3 years. It is crucial to underscore that though predicted samples in the training set reappear in the test set, their input graphs and predicted values undergo significant variations due to dynamic contextual changes. This intrinsic challenge renders the task both complex and critical. Moreover, we meticulously categorize the papers in the results for thorough and detailed analysis.

Table 3: Detailed statistics of single input heterogeneous subgraph for DPPDCC.

dataset		CS			CHM			PSY		
phase		train	val	test	train	val	test	train	val	test
nodes	paper	204	222	257	149	187	254	361	401	460
	author	446	501	613	505	646	907	1015	1173	1431
	venue	72	80	92	35	41	50	92	104	123
	time	23	24	26	22	24	27	32	34	37
edges	cites	740	827	1016	480	662	1036	1607	1808	2085
	writes	599	674	827	673	864	1218	1388	1598	1933
	publishes	165	183	216	135	168	228	311	350	408



(a) CS (b) CHM (c) PSY

Figure 1: Citation increment distribution of normal samples and focused samples cited by highly-cited papers for diffusion perspective.

In Table 3, we present detailed statistics of the input data for DPPDCC to demonstrate the sufficiency of information within multi-hop heterogeneous subgraphs. For simplicity, we average the node and edge counts by time and sample. It is evident that the Psychology dataset is the densest, while the Computer Science and Chemistry datasets are comparable.

D.2 Descriptive Statistics

To validate the rationale behind the disentangled factors used in our work, we visualize the distribution of citation increments between papers potentially influenced by corresponding popularity factors and other normal papers. For the information diffusion effect, it is posited that documents cited by highly-cited papers are more likely to attract attention from other papers during dissemination, thereby accruing more citations. Conversely, for the collective conformity effect, it is argued that the cumulative citations received by a paper can reflect its partial reputation and consequently increase its likelihood of being cited. Highly-cited papers mentioned in both contexts are classified using a Pareto distribution method: papers providing over 80% of the total cumulative citations are deemed highly-cited. The papers cited by these highly-cited papers represent focused samples of the information diffusion effect, while the highly-cited papers themselves serve as focused samples of the collective conformity effect.

(1) In Figure 1, the significant deviation among different categories is evident, with papers cited by highly-cited ones generally receiving more citations. Notably, the CS dataset exhibits the most distinct pattern, with the left peak nearly disappearing, resulting in an overall distribution closer to a normal distribution. The model aims to extract these diffusion-related factors to model the information diffusion effect independently.

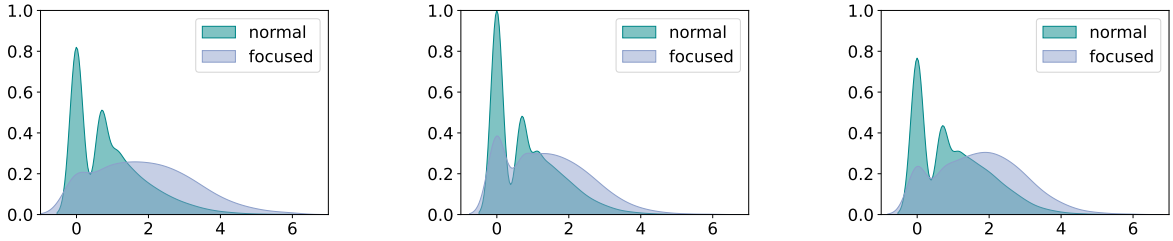
(2) Figure 2 illustrates the difference in attracting new citations between papers with accumulated citations and those with fewer citations, indicating that the collective conformity effect is more pronounced than the information diffusion effect. Notably, the distribution of focused samples in the CS dataset has essentially normalized, while other datasets also display significant peak differences. The model is expected to separately model the conformity effect, enabling it to focus more on aspects unrelated to existing cumulative citations and align more closely with contribution-related information.

D.3 Baseline Models

The baseline models are mainly divided into citation/cascade prediction models, dynamic GNN models, and disentangled GNN models:

Citation/Cascade prediction models: We apply classic and recent citation/cascade prediction models to estimate the potential impacts of papers based on content, citation graphs, or citation cascade graphs.

- **SciBERT** [2] is a BERT-based model pretrained on scientific corpus. We fine-tune it to encode paper titles and abstracts, predicting with the [CLS] token for citation prediction.



(a) CS (b) CHM (c) PSY

Figure 2: Citation increment distribution of normal samples and focused highly-cited samples for conformity perspective.

- **SPECTER2** [9] generates document-level embeddings of scientific documents by pretraining a BERT-based model on relatedness signals extracted from the citation graph. Instead of fine-tuning the model's adapters, we fully fine-tune the basic SPECTER2 for better performance.
- **HINTS** [20] tackles the cold-start problem by employing R-GCN on pseudo meta-data heterogeneous graphs. It then feeds the node (paper) representations into GRU for sequential modeling and employs a stochastic model for predictions. To integrate content information from PLMs, we replace node ID embeddings and adjust the model to encode subgraphs instead of the entire graph, mitigating memory issues.
- **CasSampling** [8] employs an importance-aware sampling method to extract critical components of cascade graphs based on outdegree, retaining the time series of the global graphs as node features. Subsequently, an attention aggregator is used to encode the global-level time flow.
- **H^{2}CGL** [8], the state-of-the-art citation prediction model, shares a similar dynamic setting and heterogeneous GNN framework with our approach. It encodes the dynamics of the citation network hierarchically and designs message-passing modules that consider the unique characteristics of citation networks. Additionally, it incorporates graph contrastive learning to refine representations, encouraging the model to encode potential citation information unrelated to specific topics.

Dynamic GNN models: We employ recent dynamic GNN models to extract paper representations from the dynamic graphs collected in the same way as our model, taking the potential impact prediction as a graph regression task. Similar to HINTS, we adapt these models to concentrate on target-centric subgraphs rather than the entire graph.

- **EvolveGCN** [33] integrates RNN structure within GNN modules. It regards graph structures as hidden states within the GRU. It adopts GRU parameters as GCN parameters.
- **Dysat** [35] incorporates multiple self-attention modules to encode structural and temporal information. It first applies a shared GAT to encode subgraphs separately across different years. Subsequently, a vanilla transformer encoder models the temporal information of the graph representation sequence.
- **ROLAND** [54] is a novel dynamic heterogeneous GCN model that inherits static GCN methods. It retains the hierarchical information of GCNs' different layers and then applies a GRU-like updater to iteratively update representations.

Disentangled GNN models: We apply recent disentangled GNN models to extract critical representations from the multi-hop subgraphs.

- **DisenGCN** [29] partitions the feature vector into K channels and uses neighborhood routing from capsule networks for disentanglement.
- **DisenHAN** [44] adapts DisenGCN to heterogeneous graphs for recommendation. It employs DisenGCNs in different homogeneous graphs for intra-aggregation. It then employs the attention mechanism to aggregate information from different relationships under various channels.
- **CAL** [40] employs attention modules to estimate the causal and trivial masks for structures and attributes. It encodes them with specific GCNs separately and applies causal intervention by randomly swapping the trivial part of the whole embeddings.
- **DisC** [12] first estimates the causal/bias masks and then employs causal/bias-aware loss functions. It further generates counterfactual unbiased samples in the embedding space by swapping the biased part of the representations.
- **DIDA** [57] discovers invariant patterns within dynamic graphs by generating soft masks using the self-attention mechanism. It then applies random variant pattern swapping in both temporal and spatial dimensions as the causal intervention.
- **DyTed** [56] designs a temporal-clips contrastive learning task along with a structural contrastive learning task to effectively disentangle time-invariant and time-varying representations. It further employs an adversarial learning framework with a disentanglement-aware discriminator for enhancement. EGCN and Dysat are utilized as the backbone dynamic GNNs of DyTed, with the best performance reported using Dysat.
- **SILD** [58] addresses distribution shifts in both dynamic and spectral domains. It builds upon DIDA with spectral transformation and proposes a disentangled spectrum mask to capture and utilize invariant and variant spectral patterns.

D.4 Implementation Details

All baselines and our proposed model are implemented using PyTorch and DGL [43]/PyG [13]. They are trained on an NVIDIA A800 80GB GPU and optimized with the Adam optimizer [22]. For the baselines, we directly use or refer to the official source codes. Experiments are conducted with three random seeds, and the averaged results are reported. Models are developed using the training and validation sets, with the final model selected based on the best main metric on the validation set. Each dataset is divided into previous papers (included in the training set), fresh papers (new additions in the test set), and immediate papers (published at the test time point) for detailed analysis. We follow the recommended settings from the original papers or open-source codes for all baselines. In our model across all datasets, the hop order k is set to 2, corresponding top limits K_1 and K_2 are set to 100 and 20, the learning rate is set to 1e-4, the number of layers L is set to 4, the co-cited/citing weight λ and the disentanglement weight β are set to 0.5, and the categories of equal frequency bins M are set to 5. Additionally, the batch size is set to 64 for CS and CHM, and 32 for PSY, and the hidden dimension d is set to 128 for CS and PSY, and 192 for CHM.

D.5 Evaluation Metric

In our evaluation, we utilize two metrics: MALE (Mean Absolute Logarithmic Error) and LogR^2 (Logarithmic R-squared). MALE calculates the absolute error between the target and predicted values post-logarithm transformation, reflecting the direct predictive capability:

$$MALE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

Regarding LogR^2, it evaluates the coefficient of determination, calculated as the proportion of the dependent variable's variation predictable from the independent variable, following logarithm transformation for both variables:

$$LogR^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \frac{1}{N}\sum_{i=1}^{N} y_i)^2} \tag{2}$$

LogR^2 assumes a negative value when the model's predictions fall short of the data's mean value, indicating an inability to capture the underlying trend within the dataset. In our problem context, mirroring practical scenarios within dynamic contexts that require extrapolative abilities, encountering negative values is both frequent and expected due to inherent complexities. Consequently, by integrating MALE and LogR^2, we holistically gauge the model's accuracy in direct predictions for individual samples and its aptness in capturing broader trends across groups.

E Complete Experimental Results

We run the experiments with three random seeds and report both the averaged results and standard deviations of all models. As shown in Figure 4, our proposed model DPPDCC consistently outperforms others with minimal variation, highlighting its capacity and stability.

F Complexity Experiment

Our model's complexity is primarily influenced by the R-GCN architecture, which requires distinct message-passing. During training, proposed contrastive learning involves additional encoding of multi-view graphs, but this step can be omitted during inference. Notably, our model converges quickly, requiring only 5 epochs compared to 20 epochs for other baselines, ensuring comparable training time. Given that we train and test on datasets with approximately 0.2 million samples, it ensures comparable coverage. Figure 5 demonstrates that our model has comparable space occupancy and time complexity to the baseline models.

G Extended Visualizations

G.1 Weighting Score Visualization

To effectively visualize the connections between distributions, we gather all scores after applying softmax based on the target node. These scores are flattened to observe the overarching trend within the latest year. The original attention scores e are averaged across all heads for ease of interpretation. Notably, we employ reversed values via $1 - c$ for the negative effect in the co-citing relationship, rendering them positive to reflect the trend. In Figure 3, robustly positive relationships are observed among the sum distribution α, original attention distribution e, and the distribution of co-cited/citing strengths c. It is reasonable to note that the sum distribution closely resembles either of the sourced distributions. Compared to the co-cited/citing distribution, the sum distribution demonstrates greater similarity to the original attention distribution. Intriguingly, within the initial layers of the model, minimal disparities are evident between the original attention scores and the co-cited/citing strengths. This phenomenon is attributed to both the averaging across all heads and the model's adaptation to the co-cited/citing information during the learning process. Moreover, while the overall trend is evident, a finer examination of the samples reveals that for instances with varying co-cited/citing strengths, their original attention scores tend to be equal, resulting in the displayed horizontal and vertical distribution evident in Figures 3a and 3e. This equal might arise due to the initial state of the attention mechanism in the GAT. However, as the model progresses in-depth, noticeable disparities emerge between these distributions in both "cites" and "is cited by" edges. In the upper layers, attention scores tend to concentrate on a select few nodes, while in the lower layers, they disperse attention

Table 4: The complete experimental results of total samples with standard deviations.

model	CS		CHM		PSY	
	MALE↓	LogR2↑	MALE↓	LogR2↑	MALE↓	LogR2↑
SciBERT	0.6904±.007	0.3370±.016	0.5868±.030	0.3887±.054	0.6131±.019	0.4761±.034
SPECTER2	0.7482±.027	0.2337±.058	0.5832±.013	0.3906±.032	0.5935±.020	0.5082±.031
HINTS	0.9022±.108	-0.1227±.231	0.8265±.074	-0.1283±.160	0.8529±.081	-0.0198±.204
CasSampling	0.5061±.003	0.6112±.007	0.4679±.002	0.6127±.003	0.5196±.002	0.6200±.006
H2CGL	0.4579±.007	0.6812±.014	0.4337±.003	0.6573±.010	0.4727±.005	0.6836±.006
EGCN	0.8051±.009	0.1296±.009	0.6381±.007	0.2826±.020	0.7339±.012	0.2530±.021
Dysat	0.6297±.006	0.4242±.015	0.5412±.009	0.4705±.017	0.6207±.047	0.4354±.096
ROLAND	0.5978±.033	0.4911±.048	0.5437±.037	0.4495±.073	0.5521±.014	0.5718±.017
DisenGCN	0.8321±.010	0.1026±.030	0.6343±.006	0.2897±.016	0.6689±.011	0.3746±.018
DisenHAN	0.8221±.006	-0.0488±.038	0.8652±.181	-0.1110±.355	0.8766±.191	-0.1186±.552
CAL	0.7094±.018	0.1859±.222	0.5733±.015	0.4166±.026	0.5535±.009	0.5783±.013
DisC	0.6733±.006	0.3496±.013	0.5609±.031	0.4259±.065	0.5592±.021	0.5606±.035
DIDA	0.6203±.034	0.4449±.045	0.5108±.016	0.5354±.035	0.5723±.022	0.5328±.034
DyTed (EGCN)	0.8101±.008	0.1235±.008	0.6411±.007	0.2465±.006	0.7896±.009	0.1478±.025
DyTed (Dysat)	0.6820±.035	0.2426±.138	0.5386±.010	0.4602±.013	0.5838±.001	0.5029±.001
SILD	0.5836±.011	0.4977±.031	0.5009±.003	0.5629±.004	0.5448±.009	0.5889±.016
DPPDCC	0.4432±.005	0.6938±.008	0.4264±.002	0.6667±.005	0.4587±.002	0.6969±.002
#improve (%)	3.21*	1.85*	1.69*	1.43*	2.96*	1.94*

Table 5: The input data, parameters, and inference time of selected baselines and DPPDCC.

Type	model	Input Data		CS		CHM		PSY	
		Dynamic	Hetero	params	inference	params	inference	params	inference
PLM	SciBERT	×	×	111.1M	2.91s	111.1M	2.92s	111.1M	2.92s
Cascade	CasSampling	×	×	0.2M	0.02s	0.2M	0.02s	0.2M	0.02s
Dynamic	Dysat	✓	×	0.2M	0.66s	0.2M	0.76s	0.2M	1.24s
	ROLAND	✓	✓	0.5M	4.76s	0.5M	6.59s	0.5M	8.71s
Disentangled	CAL	×	×	0.3M	0.07s	0.3M	0.07s	0.3M	0.11s
	DIDA	✓	×	1.4M	2.26s	1.4M	2.55s	1.4M	4.89s
	DyTed	✓	×	2.1M	0.79s	2.1M	0.78s	2.1M	1.37s
	SILD	✓	×	2.2M	1.31s	2.2M	1.42s	2.2M	2.40s
ours	DPPDCC	✓	✓	14.3M	4.29s	31.0M	6.06s	14.3M	7.67s

across a wider array of nodes. Furthermore, co-cited/citing strengths represent 2-hop information conveyed through intermediate nodes, resembling lower-hop information within GNNs. Therefore, continuously incorporating the co-cited/citing distribution across all layers reinforces signals from pivotal nodes across a broader spectrum. It can alleviate the over-smoothing issues by preventing the dominance of a select few nodes within aggregated information. This approach serves to explain why a layer count of 4 emerges as optimal—it encapsulates higher-order information beyond the 2-hop range, allowing the co-cited/citing information to complement the original attention scores. This collaboration achieves multi-hop information integration, mitigating both forgetting and overemphasis within the learning process.

G.2 Disentanglement Correlation Analysis

In Table 6, we illustrate the correlations among all disentangled perspective values/proportions and the predicted/real values, focusing on samples with all positively predicted values across all datasets.

Taking the correlation analysis of Computer Science as an example, "conformity" stands out among the values of all three perspectives. Its correlation with the predicted/real values shows a significant gap (-0.02) compared to other perspectives. Despite this, conformity emerges as the most closely associated perspective with accumulated citations, implying its significance in representing information

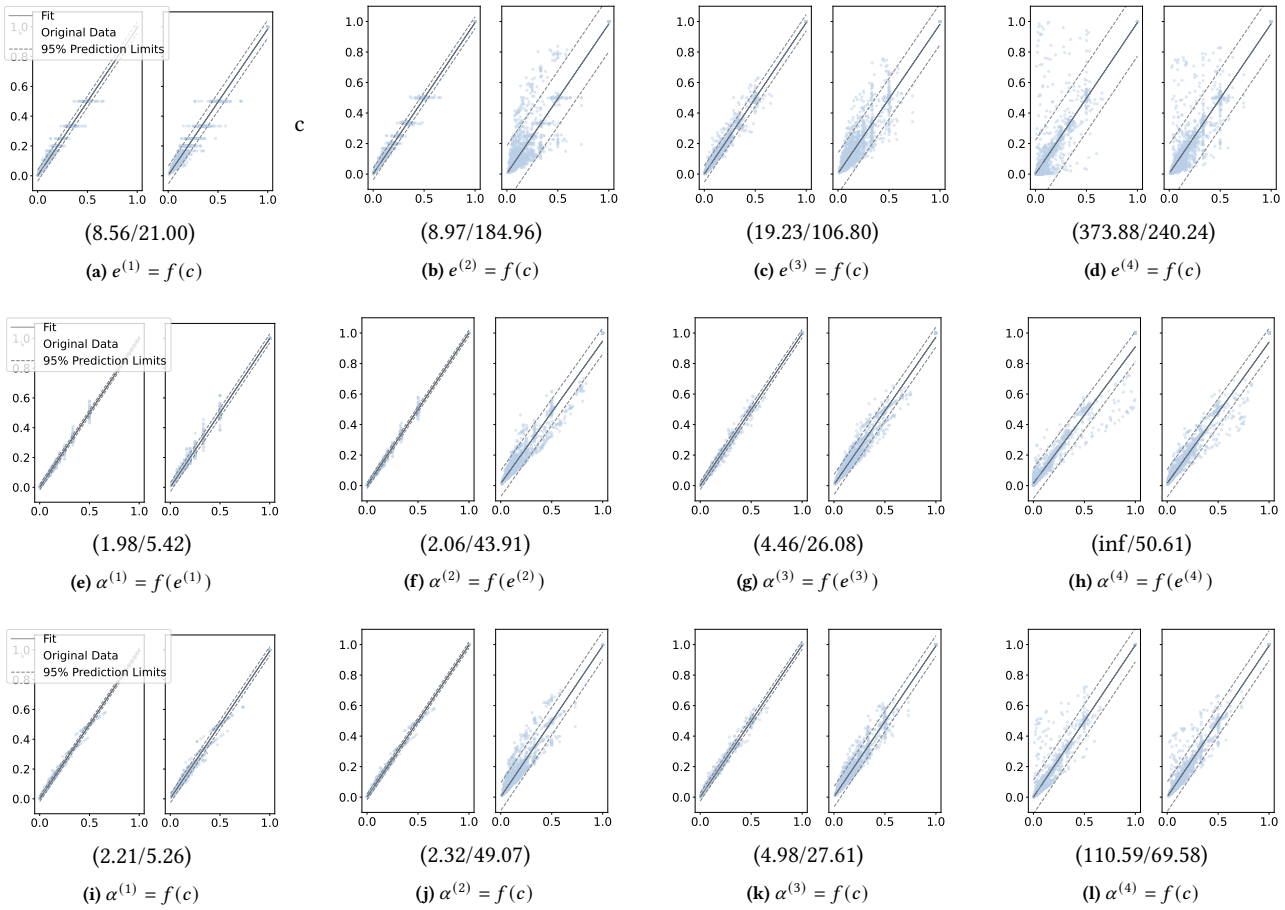


Figure 3: The outcomes of the linear regression analysis concerning the "cites"/"is cited by" scores within the field of computer science. Below the respective sub-figures, we depict the mean Kullback-Leibler (KL) divergences between the target and source distributions, denoted in units of 10^{-4}.

about a paper's reputation. The predicted increment appears as the most influential predictor of the real increment. Its aggregated value outperforms individual perspectives, highlighting the effectiveness of the proposed disentangled learning approach. Moreover, the real increment demonstrates minimal correlation with the paper's accumulated citations. This revelation underscores two key points: (1) Citation increment proves to be an effective metric in portraying a paper's potential impact, with accumulated citation accounting for only a fraction of the future increment. (2) The DPPDCC model adeptly captures the trend in a paper's citation increment, validating its substantial potential and reliability. In our calculations of the proportions of all three perspectives within predicted values, we find that conformity diffusion is the only perspective with numerous negative correlations. This suggests that in the field of Computer Science, papers characterized solely by accumulated citations tend to attract fewer new citations. Additionally, it shows an almost negative correlation with the proportion of contribution (-0.78), indicating their mutual exclusivity. Consequently, conformity and contribution have learned entirely distinct facets of citation increments. Overall, the contribution perspective is the most pivotal factor for prediction, showcasing the highest correlations in both values and proportions. However, the correlations remain a flexible threshold, where other perspectives also help determine the predicted increments. As a complement to the numerical citation increment, this approach offers a more objective evaluation, potentially aiding in identifying valuable papers by encompassing diverse and complementary information.

In summary, contribution is the most significant perspective across all datasets. Its disentangled value has the largest correlation with both the predicted and real citation increments, as does its proportion in Computer Science and Psychology. Moreover, diffusion and conformity focus on different aspects of the popularity factor. Accumulated citations consistently explain the trend of conformity values, while diffusion complements the structural influence with more flexibility regarding graph features.

Table 6: Results of Disentanglement Correlation in Computer Science

CS	dif	con	contri	dif (%)	con (%)	contri (%)	citations	final	real
dif	–	0.9579	0.9990	0.0389	-0.3237	0.4556	0.3209	0.9971	0.8276
con	0.9579	–	0.9620	-0.1176	-0.1544	0.3495	0.3575	0.9765	0.8091
contri	0.9990	0.9620	–	0.0276	-0.3236	0.4665	0.3239	0.9981	0.8287
dif (%)	0.0389	-0.1176	0.0276	–	-0.7657	0.1967	-0.0364	-0.0022	-0.0003
con (%)	-0.3237	-0.1544	-0.3236	-0.7657	–	-0.7812	-0.0056	-0.2861	-0.2477
contri (%)	0.4556	0.3495	0.4665	0.1967	-0.7812	–	0.0439	0.4382	0.3779
citations	0.3209	0.3575	0.3239	-0.0364	-0.0056	0.0439	–	0.3331	0.2982
final	0.9971	0.9765	0.9981	-0.0022	-0.2861	0.4382	0.3331	–	0.8298
real	0.8276	0.8091	0.8287	-0.0003	-0.2477	0.3779	0.2982	0.8298	–
CHM	dif	con	contri	dif (%)	con (%)	contri (%)	citations	final	real
dif	–	0.8441	0.9932	0.6134	-0.3982	-0.0478	0.1844	0.9828	0.7834
con	0.8441	–	0.8723	0.2398	-0.0062	-0.3362	0.2581	0.9267	0.7317
contri	0.9932	0.8723	–	0.5724	-0.3795	-0.0275	0.2035	0.9911	0.7912
dif (%)	0.6134	0.2398	0.5724	–	-0.9104	0.4773	-0.0080	0.5075	0.4006
con (%)	-0.3982	-0.0062	-0.3795	-0.9104	–	-0.7981	0.0477	-0.2885	-0.2331
contri (%)	-0.0478	-0.3362	-0.0275	0.4773	-0.7981	–	-0.0897	-0.1263	-0.0883
citations	0.1844	0.2581	0.2035	-0.0080	0.0477	-0.0897	–	0.2189	0.2022
final	0.9828	0.9267	0.9911	0.5075	-0.2885	-0.1263	0.2189	–	0.7955
real	0.7834	0.7317	0.7912	0.4006	-0.2331	-0.0883	0.2022	0.7955	–
PSY	dif	con	contri	dif (%)	con (%)	contri (%)	citations	final	real
dif	–	0.8642	0.9972	0.5355	-0.5871	0.6302	0.3556	0.9903	0.8196
con	0.8642	–	0.8740	0.2039	-0.2505	0.2997	0.4998	0.9244	0.7598
contri	0.9972	0.8740	–	0.5014	-0.5609	0.6149	0.3572	0.9928	0.8199
dif (%)	0.5355	0.2039	0.5014	–	-0.9869	0.9337	-0.0134	0.4522	0.3917
con (%)	-0.5871	-0.2505	-0.5609	-0.9869	–	-0.9792	0.0056	-0.5070	-0.4328
contri (%)	0.6302	0.2997	0.6149	0.9337	-0.9792	–	0.0043	0.5570	0.4683
citations	0.3556	0.4998	0.3572	-0.0134	0.0056	0.0043	–	0.4017	0.3416
final	0.9903	0.9244	0.9928	0.4522	-0.5070	0.5570	0.4017	–	0.8256
real	0.8196	0.7598	0.8199	0.3917	-0.4328	0.4683	0.3416	0.8256	–

G.3 Complete Disentanglement Composition

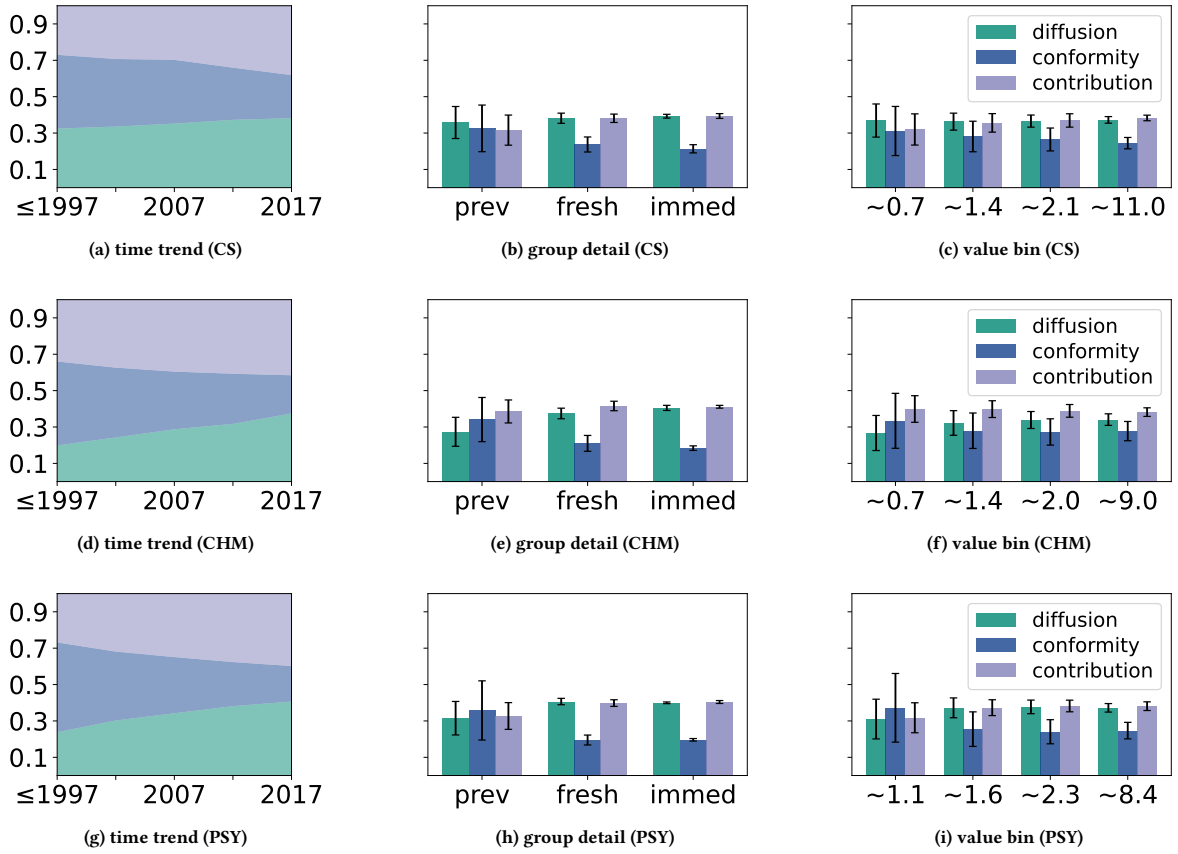
Figure 4 illustrates the disentangled compositions across all datasets. We observe common trends, such as the similar pattern of conformity and the dominance of contribution. Additionally, details such as the respective slopes and nuanced composition for various categories vary according to the intrinsic characteristics of the corresponding fields.

H Case Study

Table 7: Detailed meta-data, disentangled values/proportions, and potential impact of the selected cases.

id	meta			disentangled values			disentangled proportions			potential impact	
	pub time	refs	citations	dif	con	contri	dif	con	contri	pred	real
981400	2016	28	7 (3/2)	1.2175	0.7428	1.1862	38.69%	23.61%	37.70%	3.1466	3.0445
47429	2009	12	51 (15/3)	1.1432	0.7895	1.1993	36.50%	25.21%	38.29%	3.1320	3.1355
88781	2014	37	5 (1/0)	1.1984	0.7247	1.2307	38.00%	22.98%	39.02%	3.1538	4.7449

To delve into the intricate workings of our proposed model in addressing **RQ4**, we select three representative papers for each perspective. These papers should be co-cited or citing papers with similar topics, reflecting comparable predicted and real potential impact. Our approach involves contrasting their metadata, disentangled values, perspective proportions, and predicted/real potential impact. In this section, we visualize graph structures organized by co-cited strengths c, average original attention scores e, and sum scores α within "cites" edge subgraphs of the selected papers. To maintain clarity, we focus on graphs extracted from the latest year of the snapshots and average scores across all layers of DPPDCC. We then filter the edges in 2-hop graphs based on the top 30% accumulated weights, resulting in sparser graphs. To create target-centric connected graphs, we retain only the direct citations and references of the target paper. Highly-cited papers,



(a) time trend (CS) (b) group detail (CS) (c) value bin (CS)

(d) time trend (CHM) (e) group detail (CHM) (f) value bin (CHM)

(g) time trend (PSY) (h) group detail (PSY) (i) value bin (PSY)

Figure 4: Visualization of disentangled value proportions. (a)/(d)/(g) displays the trend that evolved with the publication time. (b)/(e)/(h) demonstrates the detailed composition of papers categorized into previous, fresh, and immediate ones. (c)/(f)/(i) is binning the samples based on the predicted values.

representing papers that provide the top 80% of citations within their respective fields of the whole datasets, are highlighted by filling in red within the graphs. Our labeling strategy encompasses the target paper and its directly connected highly-cited citations.

As outlined in Table 7, papers 981400, 47429, and 88781, respectively, embody the diffusion, conformity, and contribution perspectives. Despite their near-identical predicted increments, their disentangled value compositions differ from one another. Notably, 981400 and 88781 are recent papers not encountered during the training phase, while 47429 is an earlier publication, each with varying years since publication and distinct accumulated citations. These papers exhibit perspective proportions aligning with all trends discussed in Section E. As depicted in Figure 5, c subgraphs manifest as the densest, while e and α subgraphs appear considerably sparser, featuring fewer interconnected nodes. In the higher layers of DPPDCC, the model tends to concentrate on specific nodes, resulting in sparsity in these graphs. The introduced co-cited strengths c act as a complementary factor to the original attention scores e, facilitating the identification of significant citations from highly-cited papers. Notably, c consistently integrates significant highly-cited papers that were overlooked, replacing the original set with more valuable ones in e subgraphs. For instance, Figure 5b considers only one highly-cited paper (541312). However, in Figure 5c, one additional highly-cited paper (437859) - initially presented in Figure 5a is taken into account. Similarly, although Figure 5c and 5b share the same number of highly-cited papers (3), the co-cited subgraph c replaces one of them (37213 $\rightarrow$ 51553).

For paper 981400, within the test set, it stands as a recent publication with limited citations due to its relatively short existence. Although it has fewer accumulated citations than paper 47429, 981400 has attracted significant attention from two highly-cited papers, which is noteworthy given its total of seven citations. Consequently, its average influential citation (28.6%) exceeds that of 47429 (5.9%), resulting in a higher proportion of the diffusion perspective in its incremental source. This positions it as a crucial element in the information diffusion process, contributing to its high diffusion values. On the other hand, paper 47429, the oldest paper in the test set with a publication span of nearly 8 years by 2017, stands out with the highest conformity value among all cases, significantly contributing to a larger proportion of its incremental impact. Despite its extensive citation count, its significance has diminished over time in the evolving context of the citation network, with citations tending to emphasize its established reputation rather than other factors. Paper 88781, also recent with limited

(a) 981400-c (b) 981400-e (c) 981400-α

(d) 47429-c (e) 47429-e (f) 47429-α

(g) 88781-c (h) 88781-e (i) 88781-α

Figure 5: Weighted *paper* "cites" subgraphs of selected cases organized by co-cited strengths c, average original attention scores e, and sum scores α in Computer Science. We filter the edges within 2-hop graphs by retaining only the top 30% accumulated weights of each node and preserve the direct citation or reference nodes of the target paper to attain these sparser graphs. We only label the target paper and its directly linked highly-cited citations.

citations, lacks both highly-cited and accumulated citations compared to papers 981400 or 47429. Consequently, it exhibits low values and proportions in both diffusion and conformity perspectives. However, its contribution perspective emerges as pivotal, showcasing the largest values and proportion. Leveraging its contribution, 88781 surpasses the other two papers in predicted/real potential impact. Interestingly, DPPDCC highlights intrinsic factors to its contribution beyond popularity-related factors highlighted in graph structures or accumulated citation counts, emphasizing DPPDCC's substantial potential for practical application and impact.

Indeed, our proposed DPPDCC effectively disentangles citation increments from different factors for the target paper. This disentanglement, as evidenced by the values and proportions assigned to distinct perspectives, offers more comparable metrics for identifying valuable papers. In future endeavors, we aspire to expand this framework and evaluate its practical utility in uncovering novel insights within real-world scenarios. This continued exploration will allow us to delve further into discoveries and new applications facilitated by our disentanglement approach.

I Limitations

In our future endeavors, we intend to develop extensive datasets spanning diverse fields of study to simulate practical scenarios. This initiative is geared towards significantly enhancing the practicality and relevance of our approach. Additionally, we plan to incorporate additional factors that account for citation intent, thereby enabling a more precise evaluation of the genuine contributions made by academic papers. Furthermore, we will conduct a thorough analysis of the model's performance in identifying novel papers within specific domains. Importantly, we'll validate these findings by comparing results with relevant studies, elucidating the practical implications of our proposed methodology.

References

[1] Ali Abrishami and Sadegh Aliakbary. 2019. Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics* 13, 2 (2019), 485–499.
[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3615–3620.
[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.
[4] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. 2019. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, Vol. 2019. NIH Public Access, 2060.
[5] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1149–1158.
[6] Xueqin Chen, Fengli Zhang, Fan Zhou, and Marcello Bonsangue. 2022. Multi-scale graph capsule with influence attention for information cascades prediction. *International Journal of Intelligent Systems* 37, 3 (2022), 2584–2611.
[7] Zhihong Chen, Jiawei Wu, Chenliang Li, Jingxu Chen, Rong Xiao, and Binqiang Zhao. 2022. Co-training disentangled domain adaptation network for leveraging popularity bias in recommenders. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*. 60–69.
[8] Guixiang Cheng, Xin Yan, Shengxiang Gao, Guangyi Xu, and Xianghua Miao. 2023. CasSampling: Exploring Efficient Cascade Graph Learning for Popularity Prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 70–86.
[9] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*. Association for Computational Linguistics (ACL), 2270–2282.
[10] Emily L Denton et al. 2017. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, Vol. 30.
[11] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. 2015. Will this paper increase your h-index? Scientific impact prediction. In *Proceedings of the eighth ACM international conference on web search and data mining*. 149–158.
[12] Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. 2022. Debiasing graph neural networks via learning disentangled causal substructure. In *Advances in Neural Information Processing Systems*, Vol. 35. 24934–24946.
[13] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.
[14] Wolfgang Glänzel and András Schubert. 1995. Predictive aspects of a stochastic model for citation processes. *Information processing & management* 31, 1 (1995), 69–80.
[15] Marco Gori, Gabriele Monfardini, and Franco Scarselli. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks*, Vol. 2. 729–734.
[16] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record* 42, 2 (2013), 17–28.
[17] Guoxiu He, Zhikai Xue, Zhuoren Jiang, Yangyang Kang, Star Zhao, and Wei Lu. 2023. H2CGL: Modeling dynamics of citation network for impact prediction. *Information Processing & Management* 60, 6 (2023), 103512.
[18] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*. 2704–2710.
[19] Shengzhi Huang, Yong Huang, Yi Bu, Wei Lu, Jiajia Qian, and Dan Wang. 2022. Fine-grained citation count prediction via a transformer-based model with among-attention mechanism. *Information Processing & Management* 59, 2 (2022), 102799.
[20] Song Jiang, Bernard Koch, and Yizhou Sun. 2021. HINTS: citation time series prediction for new publications via dynamic heterogeneous information network embedding. In *Proceedings of the Web Conference 2021*. 3158–3167.
[21] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled Representation Learning for Non-Parallel Text Style Transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 424–434.
[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[23] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140* (2023).
[24] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. 2021. Learning debiased representation via disentangled feature augmentation. In *Advances in Neural Information Processing Systems*, Vol. 34. 25123–25133.
[25] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. 2017. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th international conference on World Wide Web*. 577–586.
[26] Haoyang Li, Xin Wang, Ziwei Zhang, Zehuan Yuan, Hang Li, and Wenwu Zhu. 2021. Disentangled contrastive learning on graphs. In *Advances in Neural Information Processing Systems*, Vol. 34. 21872–21884.
[27] Yanbei Liu, Xiao Wang, Shu Wu, and Zhitao Xiao. 2020. Independence promoted graph disentangled networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4916–4923.
[28] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*. PMLR, 4114–4124.
[29] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019. Disentangled graph convolutional networks. In *International conference on machine learning*. PMLR, 4212–4221.
[30] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 5711–5722.
[31] Franco Manessi, Alessandro Rozza, and Mario Manzo. 2020. Dynamic graph convolutional networks. *Pattern Recognition* 97 (2020), 107000.
[32] Marcelo Mendoza. 2021. Differences in citation patterns across areas, article types and age groups of researchers. *Publications* 9, 4 (2021), 47.
[33] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. 2020. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5363–5370.
[34] Xuanmin Ruan, Yuanyang Zhu, Jiang Li, and Ying Cheng. 2020. Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics* 14, 3 (2020), 101039.
[35] Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, and Hao Yang. 2020. Dysat: Deep neural representation learning on dynamic graphs via self-attention networks. In *Proceedings of the 13th international conference on web search and data mining*. 519–527.
[36] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.
[37] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*. Springer, 593–607.

[38] ZK Silagadze. 1997. Citations and the Zipf–Mandelbrot Law. *Complex Systems* 11 (1997), 487–499.
[39] Alessandro Sperduti and Antonina Starita. 1997. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks* 8, 3 (1997), 714–735.
[40] Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. 2022. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1696–1705.
[41] Alexandru Tatar, Marcelo Dias De Amorim, Serge Fdida, and Panayotis Antoniadis. 2014. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications* 5, 1 (2014), 1–20.
[42] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1415–1424.
[43] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv preprint arXiv:1909.01315* (2019).
[44] Yifan Wang, Suyao Tang, Yuntong Lei, Weiping Song, Sheng Wang, and Ming Zhang. 2020. Disenhan: Disentangled heterogeneous graph attention network for recommendation. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1605–1614.
[45] Qianlong Wen, Zhongyu Ouyang, Jianfei Zhang, Yiyue Qian, Yanfang Ye, and Chuxu Zhang. 2022. Disentangled dynamic heterogeneous graph learning for opioid overdose prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2009–2019.
[46] Lingfei Wu, Dashun Wang, and James A Evans. 2019. Large teams develop and small teams disrupt science and technology. *Nature* 566, 7744 (2019), 378–382.
[47] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 1 (2020), 4–24.
[48] Xovee Xu, Fan Zhou, Kunpeng Zhang, and Siyuan Liu. 2022. Ccgl: Contrastive cascade graph learning. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2022), 4539–4554.
[49] Zhikai Xue, Guoxiu He, Jiawei Liu, Zhuoren Jiang, Star Zhao, and Wei Lu. 2023. Re-examining lexical and semantic attention: Dual-view graph convolutions enhanced BERT for academic paper rating. *Information Processing & Management* 60, 2 (2023), 103216.
[50] Pengwei Yan, Yangyang Kang, Zhuoren Jiang, Kaisong Song, Tianqianjin Lin, Changlong Sun, and Xiaozhong Liu. 2024. Modeling Scholarly Collaboration and Temporal Dynamics in Citation Networks for Impact Prediction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2522–2526.
[51] Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. 1247–1252.
[52] Caipiao Yang, Peng Bao, Rong Yan, Jianian Li, and Xuanya Li. 2022. A Graph Temporal Information Learning Framework for Popularity Prediction. In *Companion Proceedings of the Web Conference 2022*. 239–242.
[53] Yiding Yang, Zunlei Feng, Mingli Song, and Xinchao Wang. 2020. Factorizable graph convolutional networks. In *Advances in Neural Information Processing Systems*, Vol. 33. 20286–20296.
[54] Jiaxuan You, Tianyu Du, and Jure Leskovec. 2022. ROLAND: graph learning framework for dynamic graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2358–2366.
[55] Tian Yu, Guang Yu, Peng-Yu Li, and Liang Wang. 2014. Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics* 101, 2 (2014), 1233–1252.
[56] Kaike Zhang, Qi Cao, Gaolin Fang, Bingbing Xu, Hongjian Zou, Huawei Shen, and Xueqi Cheng. 2023. DyTed: Disentangled Representation Learning for Discrete-time Dynamic Graph. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3309–3320.
[57] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Zhou Qin, and Wenwu Zhu. 2022. Dynamic graph neural networks under spatio-temporal distribution shift. In *Advances in Neural Information Processing Systems*.
[58] Zeyang Zhang, Xin Wang, Ziwei Zhang, Zhou Qin, Weigao Wen, Haoyang Li, Wenwu Zhu, et al. 2023. Spectral Invariant Learning for Dynamic Graphs under Distribution Shifts. In *Thirty-seventh Conference on Neural Information Processing Systems*.
[59] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*. 2980–2991.
[60] 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.