电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER THESIS

论文题目 基于图神经网络的信息传播模型和规模预测研究

学科专业	软件工程
学　　号	201821090124
作者姓名	徐　增
指导老师	钟　婷　　副教授

分类号 _____ 密级 _____

UDC 注 1 _____

学 位 论 文

基于图神经网络的信息传播模型和规模预测研究

（题名和副题名）

徐 增

（作者姓名）

指导老师 钟 婷 副教授

电子科技大学 成都

（姓名、职称、单位名称）

申请学位级别 硕士 学科专业 软件工程

提交论文日期 2021.03.15 论文答辩日期 2021.04.29

学位授予单位和日期 电子科技大学 2021 年 6 月

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

A Research of Information Diffusion Models and Popularity Prediction Using Graph Neural Networks

A Master Thesis Submitted to
University of Electronic Science and Technology of China

Discipline:	**Software Engineering**
Author:	**Xu Zeng**
Supervisor:	**Asso. Prof. Ting Zhong**
School:	**School of Information and Software Engineering**

摘　要

随着互联网和移动设备的蓬勃发展，理解和研究社交网络中的信息传播过程，在近些年得到了学界和业界的广泛关注。规模预测或流行度预测的目标是对信息进行初期观测，然后预测其在网络中传播的范围和规模。如何在复杂、快速变化、受各种内外部因素影响的社交网络中准确地预测信息级联的规模，成为该问题的主要挑战之一。大多数已有的信息级联规模预测模型依赖于人工设计的特征工程和随机过程，或者通过神经网络来对其进行端到端的学习。尽管上述模型取得了一定的成功，它们面临着许多亟待解决的重大挑战：（1）只考虑了局部结构特征，不能同时对全局和局部传播结构进行建模；（2）使用了简单的时间和结构特征建模方法，忽略了层级建模；（3）无法处理信息传播过程中的变化和不确定性；（4）无法利用无标签数据；（5）已有的数据增强方法不能直接应用到信息级联图上；（6）依赖于大量的有标签训练数据，泛化性能较低；（7）信息级联表示难以迁移到其他数据集和预测任务上。

为了解决挑战（1-3），我们提出了基于图神经网络的 CasFlow 模型，该模型对信息级联图进行非线性的层级分析并对传播过程中的变化和不确定性进行建模，它通过学习时间和结构上的级联隐藏表示来预测其规模。CasFlow 模型不依赖于特定的传播模式，它采用了变分自编码器和正则化流来同时学习节点和级联级别的潜在影响因素，其预测具有更好的准确性和鲁棒性。

为了解决挑战（4-7），我们提出了基于图对比自监督学习的 CCGL 模型，它首先在有标签和无标签数据上通过不基于特定任务的对比自监督预训练来学习信息级联图的通用表示，然后在特定下游任务上使用有标签数据来进行模型微调，最后通过专门设计的教师学生网络来对模型进行知识蒸馏和迁移学习，有效地解决了“负面迁移”问题。CCGL 模型通过模拟信息在网络中的传播过程，创新性地设计了图数据增强策略 AugSIM，缓解了模型在小数据集上训练所导致的过拟合现象，并且具备更好的泛化性能。CCGL 模型可以从数据中学习到通用的知识，并将其迁移到其他类型的数据集和预测任务上以提升预测效果。它的“无监督预训练、模型微调、知识蒸馏”范式，对信息级联预测模型的设计提供了新的视角。

本文在多个公开的大规模信息级联数据集上进行了大量的实验验证，与多个常见的基准模型相比，本文提出的两个模型均显著地降低了预测误差。

关键词：信息传播，信息级联，规模预测，图神经网络，社交网络分析

ABSTRACT

With the rapid development of the Internet and mobile devices, understanding in-network information diffusion in social networks attracts much attention in both academia and industry, becoming a fundamental research problem in many real-world applications of social network analysis. Popularity prediction aims to predict the final diffusion range or size after observing the early-stage evolution of the information cascade. How to accurately predict the popularity of the information cascade in complex, rapidly evolved, and vulnerable to be influenced information network becomes one of the most challenging problems in this field. Most of the existing prediction models rely either on hypothesized stochastic processes, artificially designed feature engineering, or end-to-end deep neural networks. These models achieved some extent of success in information cascade popularity prediction, however, facing several notable challenges: (1) only consider the local structures of information cascades, they cannot simultaneously capture the underlying global and local structures; (2) use simple temporal or structural modeling techniques, they cannot jointly model these characteristics in hierarchy; (3) cannot capture the variations and uncertainties in information diffusion; (4) they cannot utilize unlabeled information cascade data; (5) existing data augmentation techniques cannot be implemented in information cascade graphs; (6) they rely on massive labeled data, which generalizes poorly; (7) the learned information cascade representations cannot be transferred across different information cascade datasets and prediction tasks.

In order to address challenges (1-3), we propose CasFlow, which is a probabilistic popularity prediction framework based on graph neural network and variational inference. It conducts non-linear hierarchical analysis on information cascade graphs and models the variations and uncertainties of information diffusion in social networks. CasFlow allows efficient information diffusion inference and models the diffusion process by learning the latent representations of both temporal and structural characteristics of information cascade graphs. CasFlow is a pattern-agnostic model leveraging the variational auto-encoders and normalizing flows to learn node-level and cascade-level latent influence factors and uncertainties. CasFlow has a better prediction performance and robustness.

In addition, to address the challenges (4-7), we propose CCGL, which is a prediction model based on graph contrastive self-supervised learning. In particular, it first learns

general representations of information cascade graphs by a task-agnostic contrastive self-supervised pre-training on both labeled and unlabeled data. Then it fine-tunes the model by utilizing labeled data in a task-specific manner. It finally uses a teacher-student network for knowledge distilling and transfer learning, which effectively address the "negative transfer" issue. CCGL model simulates the diffusion process of information in social networks and designs a novel data augmentation strategy AugSIM for information cascade graph. CCGL can effectively alleviate the overfitting problem when training on small datasets, and possesses a better generalization capability. CCGL can learn general knowledge from information cascade data, the learned knowledge can be transferred to other information cascade datasets and prediction tasks for performance improvements. The "unsupervised pre-training, fine-tuning, and knowledge distilling" paradigm of CCGL is a promising direction for the design of future information cascade prediction models.

At last, this thesis conducts extensive experimental evaluations in several public large-scaled information cascade datasets. Compared to state-of-the-art baselines, both of the proposed prediction models decreased the prediction errors significantly.

Keywords: Information diffusion, information cascade, popularity prediction, graph neural network, social network analysis

目 录

第一章 绪 论

1.1 研究工作的背景与意义

近些年来随着无线通信技术、互联网、微型设备及移动设备的快速发展和普及，人们获取数据和信息的途径，与他人进行交流和交互的方法，发生了极大的改变。理解信息的传播过程，以及信息在传播的过程中哪些因子占据主要因素，是当今社交网络分析领域中的一个研究热点。如何准确地预测信息传播的内在机制和规模，是许多现实应用中非常关键和有挑战性的任务之一。信息传播模型广泛地使用在市场营销[1]、广告投放[2]、科学影响力预测[3]、社交网络推荐系统[4]、竞选策略[5]和流行病预防[6]等重要应用中。

信息传播的轨迹和网络结构，以及信息传播过程中的参与者和传播者的序列，构成了信息级联。许多科学家和商业公司开始研究如何对各种信息传播网络进行学习和建模，例如蜂窝网络、在线社交网络、论文引文网络、内容共享网络等。对信息传播的研究，尤其是预测信息或者信息级联的规模，吸引了学界和业界的广泛关注。信息规模预测可以从多个角度进行定义，例如，它可以对社交媒体中微博或者话题标签的流行度进行预测[7–11]（微博、Twitter、Facebook、豆瓣等）；对用户分享的图片或者视频的点赞量进行预测[12]（Instagram、微信朋友圈、小红书等）；对音乐或视频的访问量进行预测[13–15]（QQ 音乐、网易云音乐、腾讯视频、YouTube 等）；对电影的票房或者评分进行预测[16]；对科研论文的引用数量进行预测[3]；对知识问答社区中的回答支持数进行预测[17]（知乎、Stack Overflow、Quora 等）；对新闻的访问量和评论数进行预测[18]；对用户的社交影响力进行预测[19]；等等。研究者们提出了许多信息级联规模预测模型，它们可以被分为以下的三种主要的类型：

（1） 基于特征工程的模型：这一类模型[7,10]主要关注如何对信息级联中的各种有利于规模预测的特征进行挖掘和建模。这些模型要求建模者具有很强的领域知识，它们的泛化性能一般较弱，某些适用于特定领域的特征难以迁移到其他信息领域。并且，许多有关用户的特征由于个人隐私政策而难以获取，例如用户主页、社交关系、用户个人的社交信息等。

（2） 基于概率的生成模型：这一类模型[10,20]对信息的传播过程进行概率学习，例如对各种信息转发的到达过程强度函数（Arrival process intensity function）进行建模。这一类模型有着严格的数学推导，并且其预测的可解释性强。但是，它们一般要求对信息级联进行长时间的观测，也不能完全

利用传播过程中所产生的各种信息，从而导致预测效果欠佳。

（3） 基于深度学习的模型：得益于计算机算力的发展，基于神经网络的深度学习技术在近些年产生了极大的进展，在许多复杂任务上取得了非常好的效果。研究者们[21,22]使用了各种深度学习技术来构建信息传播模型。循环神经网络和注意力机制被用来对信息传播中的时序或序列过程进行学习；图神经网络被用来对社交网络中的图结构数据进行建模。不过，当前的模型难以有效率地对大规模图数据进行结构学习，并且它们没有考虑到信息传播和节点嵌入中的不确定性。

上面提及的传统模型在信息级联规模预测中取得了一定成功，但是，它们在实际预测中面临着多个关键的困难和挑战，限制了它们的预测效果：

（1） 高效的信息级联表示难以获取。因为信息级联的规模分布极度扭曲，部分信息级联的规模可达数百万之多，这使得许多图嵌入模型（例如基于随机游走和图卷积神经网络的模型）难以对其进行高效的建模。已有的预测模型无法对信息级联同时进行全局和局部结构建模。对于一个大规模的社交网络来说，对其进行完整的嵌入学习需要消耗非常大的计算资源，甚至完全无法对其进行嵌入学习。

（2） 忽略了对信息级联的时间和结构特征进行同时建模。信息的初始传播阶段对预测其最终的规模至关重要。但是在实际的预测中，我们通常难以观测到完整的结构信息。如何在一个信息受限的条件下捕捉信息潜在的结构特征，是高效的规模预测模型中非常关键的一环。另外，时序特征，例如信息级联参与者的先后顺序和到达时间，信息级联的传播速度等，对准确的规模预测起到不可忽视的作用。

（3） 缺少对信息级联特征的层级建模。在传统的信息级联规模预测模型中，有些根据少量的观测来对信息传播规模进行粗略的估计，有些对用户级别的信息进行建模（例如用户会不会参与到信息级联的传播过程中）。它们没有对信息级联的节点层级（低阶）和级联层级（高阶）的特征进行层级建模。

（4） 缺少对信息级联传播过程中的变化和不确定性进行建模。理解信息传播中的变化和不确定性对信息级联的规模预测是非常重要的，例如，观测到的信息传播过程天生地包含有噪音数据，其之后的传播过程也因为各种内部和外部的影响而面临着极大的不确定性[12]。已有的预测模型并没有对传播不确定性进行建模。

（5） 传统的信息级联预测模型依赖于大量的有标签数据，并且无法利用无标签数据。这些模型通常需要大量的有标签数据来进行监督学习，但是在实际

任务中，有标签数据可能因为各种隐私政策而难以获取，或者获取的成本过高。而且，监督学习模型无法从大量存在的无标签数据中进行知识学习和建模，这导致它们的泛化性能较弱，而且容易导致严重的过拟合现象。

（6） 已有的数据增强策略也无法直接应用到信息级联图上。传统的自然语言处理或者计算机视觉领域的数据增强策略往往基于特定的数据结构，例如文本和图片等。它们所使用的数据增强策略，包括旋转、预测颜色、剪切、高斯模糊等方法，无法直接利用到信息级联图上。

（7） 基于特定任务训练的监督学习模型往往只适用于特定的数据和任务，无法将学习到的知识有效地迁移到其他类型的数据集和预测任务上。

如何解决上述难题对于我们理解信息传播的过程以及提高信息级联规模预测的效果具有至关重要的意义。在本文的之后部分，我们通过介绍相关工作来帮助读者更好地理解信息级联规模预测问题，并且指出已有模型所面临的局限性。然后我们介绍本文中提出的预测模型是如何解决上述困难和挑战的。

1.2 信息级联规模预测的国内外相关研究历史与现状

信息级联的建模和预测在近些年引起了学界和业界的广泛关注[23]。由于信息其本身天然具备的多样性和广阔的研究范围，已有的工作聚焦于不同类型的信息传播过程。在本章中，我们详细阐述已有的信息级联规模预测文献，并将其大致分为三种类型。

1.2.1 基于特征工程的预测模型

对信息中的特征进行抽取是信息级联规模预测中普遍应用的一种模型，它们可以被分为分类（Classification）或回归（Regression）任务，提前预测（Prior）或之后（Posterior）预测，微观（Micro）预测或宏观（Macro）预测等。基于特征工程的预测模型通常会利用各种机器学习技术来构建、提取、挑选各类对预测有用的特征，例如，解释网络中的传播模式，分析时间序列的发展趋势，构建预测模型等。在本节中我们介绍常见的特征种类和预测模型，以及基于特征工程的预测模型的优缺点。在本文中我们将信息级联特征分为四类：时序特征、结构特征、用户特征，以及内容特征。

1.2.1.1 时间特征

时间特征被认为是信息级联规模预测中最重要的特征之一[12,24]。常见的时间特征包括：

（1） 观察时间

时间特征通常来自于对信息级联早期的观察（例如，对信息级联的早期参与者观察一段固定的时间，或者观察一定数量的早期参与者），观察所得到的时间序列可以被用来进行特征抽取和选择。因为时间序列的长度是高度不规整的，例如，在一段固定的时间内，一些信息级联的参与者数量可以达到数千甚至数万，而大多数的信息级联只会得到很少的关注，直接将时间序列当作特征在实际应用中并不可行。对时间序列进行计算通常需要对其进行预先的转换[25]，例如，将其分割为平均分布的间隔时间段，或者只观察固定数量的时间序列。早期的预测模型[26]将观察时间结束时的信息级联规模当作特征来预测其最终规模。它们发现早期规模和未来规模之间有着很强的相关性，从而把早期的信息级联规模当作特征输入到线性预测模型之中来预测最终的信息级联规模。

（2） 发布时间

时序特征中另外一个重要的特征是发布时间。之前的研究工作[23,26–29]指出，信息级联的最终规模与其发布的时间有着非常大的关联，例如，相比发布在午夜的信息，发布在白天的信息更有可能变得流行（尽管它们之间的竞争更大）。为了缓解规模预测中用户活动时间的影响，研究者们提出了许多解决方案。例如，在 [30] 中，作者设计了 24 种不同的局部模型，每个模型分别对应自然天中的一个小时，该模型由在这一个小时内发布的数据来训练。在 [31] 中，作者设计了一种新的特征叫做推特时间（Tweet time），来消除不平衡的用户昼夜活动（Diurnal activity）影响。还有其他种类的时间特征被用来提升模型的鲁棒性，例如 Digg 时间[26]、Source 时间[29]、用户活力可变性（User activeness variability）[32] 等。

（3） 首个参与时间

信息级联的第一个参与者到达的时间也被认为是非常重要的特征之一。衍生的时序特征还包括平均到达时间、平均反应时间、到达速度变化率、休眠期、高峰时间占比等。

（4） 增长趋势

对信息级联的增长趋势进行建模被研究者们认为是规模预测中非常有价值的方法[25,33,34]。时间序列增长的模式可以被分为多种类型，例如，平缓增长型、突然爆发增长/衰减型等。如图1-1所示，在 APS 数据集上，论文引用信息级联的 20 年增长趋势被分为 10 种类型（使用 agglomerative 层级聚类算法[35] 来进行聚类）。

尽管时序特征被认为是最重要的特征之一，最近的研究显示，它们在某些情

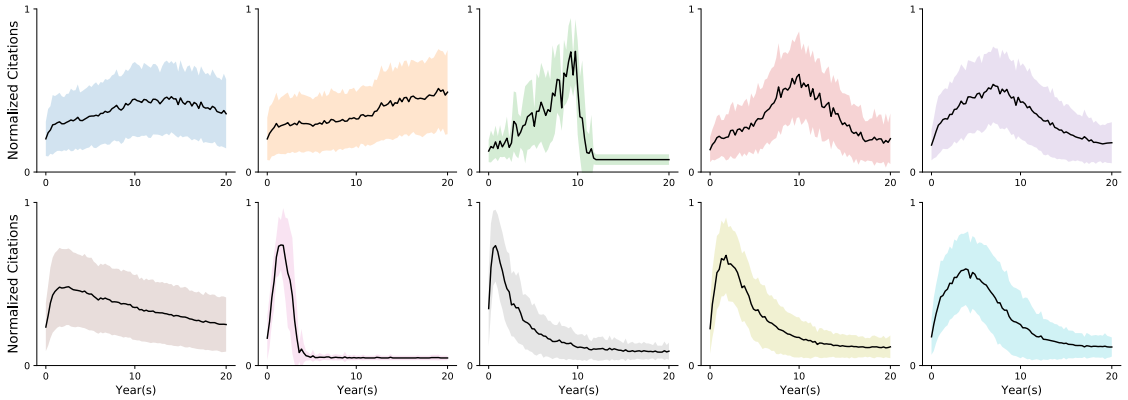


图 1-1 APS 数据集中论文信息级联的 10 种增长趋势示意图

况下也许表现的并不好。例如，它们的重要性随着时间的增长而减弱[36]，它们并不总是比其他类型的特征要好[37]，等等。

1.2.1.2 结构特征

信息级联的结构是信息传播研究中的重点。相关工作可以从它们研究信息级联的不同角度来分类：（i）只有信息级联内部的参与者组成，即信息级联图（Information cascade graph）；（ii）全局图（Global Graph），其中包括整个社交网络的所有用户，既有参与者，也有非参与者；（iii）r 阶可达图（r-reachable graph），其中包含了信息级联图的周边结构信息，例如，一阶可达图即是信息级联图在全局图中的邻居图。

信息级联图描述了信息在网络中的传播过程，其中的传播方向和传播结构包含着重要的信息。早期研究信息级联图结构的工作[38]利用了信息级联图结构来进行规模预测。例如在 [39] 中，作者们设计了两种结构特征：图边缘密度（Edge density）和图深度（Depth）。他们发现低的图边缘密度和高的图深度预示着信息级联的早期参与者们的关系更多样化，信息级联图的规模也更有可能发展的更大。

其他类型的信息级联图特征包括：节点度（Node degree）、特征向量（Eigenvector）、亲密度（Closeness）、中间性（Betweenness）、中心性（Centrality）、分值（Authority/Hub score）、图的深度（Depth）、密度（Density）、直径（Diameter）、结构多样性/扩散性（Structural diversity/virality）等。

1.2.1.3 用户特征

时间和结构特征需要对信息级联的传播过程进行早期观察，这一前置要求有时候并不现实。而用户特征可以在信息进行传播之前就可以被利用，从而使预测模型可以在信息发布之前或发布时就进行规模预测。

用户行为在信息的传播和交互（例如，用户的观看、关注、评论、分享等行为）中发挥着重要的作用。粉丝数（Number of followers）是用户特征中最为重要的特征之一，它代表着用户的社交影响力，隐含地预示了信息传播的速度和未来的规模大小[40]。拥有巨大粉丝量的用户，例如名人、新闻机构等，他们的信息更有可能传播的更远，因为他们在网络中的可见度比普通用户更大[41–43]。不过，大的信息级联并不总是由有影响力的用户所产生的。研究普通人发布的大规模信息级联的结构也很有意义[44]。

还有许多用户特征被已有的预测模型使用，例如，用户的主页（昵称、年龄、地区、教育、工作、账户创建时间）[45]、历史行为（发布内容的频率、与其他用户的交互、在线时间）[41,43,46]、用户兴趣[47]、相似度[48]等。尽管用户特征对信息级联规模的大小有着非常重要的影响，但是对于个体信息来说，宏观的用户特征无法对它们进行有差异化的预测。

1.2.1.4 内容特征

信息中所包含的内容被认为是推动信息传播的重要内在因素之一，例如，重大新闻、谣言、假信息、热点话题、矛盾话题等内容，要比普通的信息更容易获得人们的关注。

信息的文本内容是已有的规模预测模型建模的重点。文字在信息中无处不在，广泛的存在于各种文章、博客、甚至可以从音频、图片和视频中提取。已有的规模预测模型使用各种语言模型来对用户生成内容（User-Generated Content, UGC）进行建模，例如广泛使用的 Term Frequency-Inverse Document Frequency (TF-IDF) 和 Latent Dirichlet Allocation (LDA) 模型[49]。这些模型从文本中抽取特征，然后利用机器学习技术（例如朴素贝叶斯和线性回归等模型）对信息进行规模预测。在 [46] 中，TF-IDF 和 LDA 被用来学习推特中话题的分布。在 [47] 中，TF-IDF 被用来计算用户发表的推特中关键词的重要性，然后计算用户历史发表的所有推特和当前被查询的信息之间的互相关性（Mutual correlation），从而判断当前用户有多大的可能参与到被查询信息的传播之中。在 [50] 中，作者们分析了 Digg 社区中评论文本的语义和概率特征，例如评论的长度、动词或名次的数量、内容熵值（Content entropy）、可读性、主观性或客观性等特征，他们发现，用户更喜欢转发短的、简单的、可读性高的内容。

对图像的内容进行提取和建模与文本相比有很大的不同。研究者们利用计算机视觉领域的技术来对图像进行分析。一些基础的特征可以很容易地从图像中抽取，例如，在 [51] 中，作者们分析了图像的基础属性，它们包括图像的大小、时间、日期、季节、方向（水平/垂直）、设备、主导颜色、闪光灯、分辨率、地理位

置、图像的描述、标签等属性和特征。

在一篇关键的论文[52]中，作者们分析了 Flicker 图像的特征与它们流行度（例如访问量）之间的关联。他们将图像的内容特征分类为简单的人类可解释特征、低阶图像特征和高阶图像特征。这些特征被证明可以显著地提升预测效果，例如，拥有引人注目颜色的图像更容易变得流行。低阶的图像特征包括要点（Gist）、纹理（Texture）、色块（Color patches）、梯度（Gradients）等。高阶的图像特征是从 ImageNet 分类器[53]中抽取的 1,000 个分类。其他类型的内容特征还包括从视频中抽取的特征，例如视频长度、分辨率、帧数等。

1.2.1.5 预测模型

基于特征工程预测模型的主要挑战在于如何设计特征以及选择何种特征来设计模型，具体的预测模型并不是其关注的重点对象。例如，在 [12] 中，作者们发现大多数的机器学习模型，除了时间复杂度和空间复杂度外，它们的效果差异并不显著。

1.2.1.6 基于特征工程预测模型的优缺点

基于特征工程的预测模型与其他类型的模型相比，一般被认为预测效果具有竞争力，可解释性强[54]。但是，它们在现实应用中的局限性在于其依赖人工设计的特征工程。一些特征由于用户隐私政策而难以获取，例如用户偏好和历史活动等。另外一些特征则难以计算或者计算消耗过大，限制了模型的可用性。一般来说，大多数的时间特征和用户特征比较易于计算和抽取。而结构特征，尤其是对于大规模的社交网络，计算结构特征需要消耗大量的计算力。基于内容的特征，根据不同的问题定义和数据来源，拥有不同的时间和空间复杂度。

在给定一系列定义好的特征集合，如何选择数量相对较少的特征集合，同时还具有较好的预测效果，成为特征工程模型的一大挑战。已有的模型经常要求一些难以获得的特征，并且，它们的泛化性能受到很大的限制，某些特征在特定的平台和任务上表现的很好，但是在其他平台或任务上，其预测效果可能会变得非常差。因此，将其迁移到其他平台或任务上，需要重新进行费力的特征工程。

1.2.2 基于概率的生成模型

许多现实生活中的现象，例如，信息转发、医院挂号、论文引用等，可以被构建成连续时间域中的事件序列。对事件序列或者用户参与信息级联的时间序列进行建模是理解信息传播内在驱动因素的重要方法之一。许多预测模型用基于概率和统计的生成模型（Generative models）来对信息传播过程进行刻画和建模，例

如流行病模型（Epidemic models）、生存分析（Survival analysis）以及一系列随机点过程（Stochastic point processes）等。在本节中，我们对信息级联规模预测的生成模型进行介绍。

1.2.2.1 柏松过程

基于点过程（Point processes）的模型与基于特征工程的模型不同，它们从统计的、概率的、生成式的角度对信息级联进行建模。点过程经常被用来对时间序列进行分析，例如，顾客的到达率、电话拨打率、机械故障率等，在排队论和运筹学中得到了广泛应用。

在 [3,55] 中，作者们提出了一个基于加强柏松过程（Reinforced Poisson process）的生成概率模型 RPP，其被用来预测信息的规模。该模型具有三个关键的部分：（i）信息的吸引率；（ii）时间衰减函数；（iii）强化机制。

基于 RPP 模型，在 [56] 中，作者们提出了一个扩展模型 PETM。与 RPP 不同（预测长期科学影响力），PETM 被设计为预测社交领域中的信息规模。PETM 将 RPP 中的 log-normal 分布替换为幂律分布（Power-law distribution），因为幂律分布更符合社交网络中的特点。RPP 模型中的线性强化机制被替换为指数衰减强化函数（Exponential decay reinforcement function），从而可以对大规模的信息进行调整。

1.2.2.2 生存分析

统计学中的生存分析（Survival analysis）在各种领域（工程、经济、社会学等）中广泛使用。在 [57,58] 中，生存分析的概念被用来预测在线内容的规模。作者们提出了基于 Cox proportional 风险回归 [59] 模型，其具有两个主要模块：（i）一系列可解释的风险因子；（ii）基于威布尔分布的基础函数。

在 [60] 中，作者们提出了一个动态的自我模型（Dynamic egocentric model）来对引用网络进行建模，其基于一种计数过程（Counting process）。生存分析中的威布尔分布（Welbull distribution）也被用来对信息级联进行建模 [61]。作者们提出了 NEWER 模型来描述微博全局网络中的信息传播。

1.2.2.3 自激励霍克斯点过程和传染模型

基于自激励点过程的模型把事件的到达概率作为和时间及历史事件有关的函数。在 [62] 中，作者们用两个关键因素来描述 YouTube 视频的观看行为动态：（i）一个回应函数，表示人类行为等待时间的幂律分布；（ii）一个传染过程，其可以被自激励霍克斯条件柏松过程（Self-exciting Hawkes conditional Poisson process）所建模。

在 SpikdeM 模型[28] 中，作者们对信息传播中的指数增长和幂律衰减现象进行建模。他们把传染模型和自激励霍克斯过程的优点加以利用。

双重情绪霍克斯过程（Dual sentimental Hawkes process, DSHP）是另外一种基于霍克斯过程的生成模型[63]。DSHP 考虑了信息中蕴含的情绪因素，并且提出了两种自激励的点过程：自激励和交叉激励（Cross-excitation）。

自激励信息级联（Self-exciting model of information cascades, SEISMIC）模型[10] 被用来预测社交网络中的转发数量。它用一个记忆核函数（memory kernel function）来对人类反应时间进行建模，该函数符合幂律分布。SEISMIC 还对推特的转发能力（Retweetability）进行了建模。

1.2.2.4 生成模型的优缺点

生成模型一般来说不需要繁重的特征工程，可解释性也很强。它们主要依赖于时间数据，它们的预测一般来说也很效率，可以进行实时预测。生成模型的主要缺点来自于对它们预测效果的批评[21,54]，它们也很容易被离群数据所影响。另外，生成模型通常依赖于对已有机制和参数的强烈假设，这导致它们的泛化性能和建模能力较弱。对信息传播机制的低估和简单化，也让它们的预测效果不尽人意。大多数的生成模型也缺少对网络结构的建模。

1.2.3 基于深度学习的预测模型

近些年来，随着计算机算力的提高，神经网络再次得到了广泛的关注，基于深度学习的预测模型在许多任务上获得了成功。深层神经网络被认为比线性模型更为强大。例如，基于循环神经网络的预测模型并不依赖于显式的对于信息传播的机制假设，从而在捕捉信息级联时序长依赖的情况下更为灵活。基于图表示学习的预测模型不需要工程人员人工设计耗费时力的信息级联结构特征，例如特定的节点影响力和社团检测（Community detection）。

已有的基于深度学习的模型可以被大致分为三类：（i）基于信息内容的模型，例如文本、图像、视频等多媒体内容，这些模型通常使用来自于计算机视觉和自然语言处理领域的技术来对信息的内容学习有效的表示；（ii）基于时间序列的模型，这些模型对社交网络中的信息级联进行线性建模，依赖于循环神经网络、池化机制（Pooling mechanism）、注意力机制（Attention mechanism）[64] 等技术；（iii）基于图结构的模型，例如信息级联图或者全局图等，这些模型通常利用图神经网络和图表示学习技术来学习节点、边、图的有效结构表示。其他的深度学习技术，例如变分推断（Variational inference）、强化学习（Reinforcement learning）等技术也被使用在信息级联规模预测之中。在许多情况下，多模态（Multi-modalities）、

多尺度（Multi-scale）、多任务（Multi-task）学习等技术被用来提升预测的效果。

DeepCas[65]是第一个使用图表示学习技术来建模和预测信息级联规模的模型。它借鉴了 DeepWalk 模型[66]的思想，使用随机游走方法来对信息级联图进行采样。采样的节点序列被输入到双向门控循环单元[67]中，再配合注意力机制[68]来获取节点的嵌入。DeepCas 模型的预测是端到端的，从而不依赖于人工的特征设计。随后，在 [69] 中，作者们提出了 DCGT 模型，它在 DeepCas 的基础上增加了对节点内容的建模。

DeepHawkes 模型[21]的目的是将生成模型的优点和深度学习技术的优点结合起来，从而兼顾预测可解释性和良好的预测效果。DeepHawkes 模型结合了三种来自霍克斯过程的关键机制：（i）用户影响力；（ii）自激励机制；（iii）时间衰减影响。与 DeepCas 相似，DeepHawkes 用端到端的方式学习用户的表示。但是，与直接对信息级联图进行结构建模相比，DeepHawkes 使用门控循环单元、和池化（Sum pooling）、以及非参数化的时间核函数来对信息级联中单个传播路径中的早期参与者进行聚合。

ANPP 模型[70]使用了 GloVe[71]来对信息内容的文本进行嵌入，并且利用了 node2vec[72]来对用户图进行嵌入。ANPP 使用了注意力机制来对获取的表示以及时间序列特征向量进行聚合。

DTCN 模型[73]通过学习用户和图像的嵌入，分享序列的时间上下文，以及多步时间注意力机制，来预测 Flickr 图像的流行度。DTCN 模型使用了 ResNet[74]和长短期记忆人工神经网络[75]来对图片的视觉依赖和时间依赖分别进行建模。

循环级联卷积网络（Recurrent Cascades Convolutional Network, CasCN）[76]将信息级联图当作一系列子信息级联图，然后使用一种动态多方向的图卷积网络来学习信息级联的结构信息。

Coupled-GNN 模型[77]利用了两个特别设计的图神经网络来捕捉信息中的级联效应：（i）一个图神经网络对节点状态进行建模；（ii）另外一个图神经网络对影响扩散进行建模。

变分级联图学习神经网络（Variational Cascade Graph Learning Neural Network, VaCas）模型[78]结合了图小波（Graph wavelets）、变分自编码器（Variational auto-encoders）、以及双向门控循环网络（Bi-GRU）来学习级联图的结构表示。VaCas 同时对节点级别和级联级别的传播不确定性进行建模，还包括了基于上下文的用户行为表示。

一般来说，基于深度学习的模型通过各种深度学习技术来学习信息级联不同方面的表示[79]，例如，捕获时间序列之间的长时间依赖关系、使用循环神经网络

及其变种来对信息级联的参与者之间的时序特征进行建模、通过深层语言模型和视觉模型来学习级联中的文本表示和图像表示、使用无监督或半监督的图嵌入模型来学习信息级联的结构表示等。

与基于特征工程的模型或者生成模型相比，基于深度学习的模型一般不需要繁重的人工特征工程（依赖于特定平台或专家知识），也无需对信息传播的内在机制做出假设（缺少灵活性，亦依赖于特定的设计）。

尽管基于深度学习的模型在信息级联预测任务上取得了很好的效果，它们也面临着很多限制和挑战。深度学习模型的一大缺陷是缺少对其预测的可解释性，因为神经网络天生就是一种“黑盒模型”。深度学习模型的计算消耗一般也比其他两类模型要大。为了取得令人满意的预测效果，工程人员往往需要对其进行复杂的参数调整、模型训练，以及面临着对数据过拟合的风险。

1.3 本文的主要研究内容

本文在节 1.1 中指出了信息级联规模预测的七大挑战，即：（1）只考虑了局部结构特征，不能同时对全局和局部传播结构进行建模；（2）使用了简单的时间和结构特征建模方法，忽略了层级建模；（3）无法处理信息传播过程中的变化和不确定性；（4）无法利用无标签数据；（5）已有的数据增强方法不能直接应用到信息级联图上；（6）依赖于大量的有标签数据，泛化性能较低；（7）基于特定任务训练的监督学习模型往往只适用于特定的数据和任务，无法将学习到的知识有效地迁移到其他类型的数据集和预测任务上。为了解决上述挑战，本文设计了两个创新的预测模型 CasFlow 和 CCGL。

针对挑战（1-3），我们设计了基于图神经网络架构的 CasFlow 模型。它可以对全局和局部的信息传播结构进行层级建模，同时考虑到了信息级联中的时间特征。CasFlow 模型提出了以下方案来解决已有的挑战：使用图小波（Graph wavelets）来学习信息级联的局部表示，同时支持对不同大小的传播图进行建模；使用稀疏矩阵分解（Sparse matrix factorization）来学习全局的用户表示，学习到的表示可以有效地对社交网络中的用户行为和交互进行建模；构建了一个新颖的语境化传播嵌入模块来学习复杂的同一用户在不同环境下的行为；为了更好地理解用户层级的行为和级联层级的传播效应，CasFlow 构建了一种基于概率隐藏变量的层级变分自编码器（Hierarchical variational auto-encoder）来学习细粒度的信息传播结构模式；CasFlow 使用了基于变分推断（Variational inference）和正则化流（Normalizing flows）的生成模型，使其可以学习到可解释的和灵活的表示，这些表示中蕴含了信息级联中节点之间的复杂分布和长时间依赖，从而可以对图中每

一个节点的行为不确定性进行建模，以此来提升信息级联规模预测的效果。

针对挑战（4-7），我们设计了基于图对比自监督学习的 CCGL 模型。它解决了将对比自监督学习应用到信息级联图数据上的四大困难：如何用一种对比、自监督、以及不基于特定任务的方法来从信息级联图中学习到通用的知识；如何在对比学习框架下设计信息级联的正负样本并同时捕获到数据中的变化和信息级联的动态传播特性；如何对预训练的模型进行针对于下游任务的半监督微调（Semi-supervised fine tuning）；如何对模型进行知识蒸馏（Knowledge distillation），在多个数据集和预测任务中学习到通用的知识，从而可以缓解“消极迁移（Negative transfer）”问题。CCGL 模型基于半监督学习框架来对信息级联图进行对比表示学习。它首先对信息级联图进行图数据增强（Graph data augmentation）：手动地执行图扰动（例如增加和删除图中的节点和边，更改节点属性等），从而对图中的信息传播进行模拟。CCGL 在不基于特定任务的预训练阶段同时利用了有标签数据和无标签数据，在基于特定任务的模型微调和蒸馏阶段对学习到的知识进行迁移。

1.4 本文的主要创新与贡献

信息传播模型和信息级联规模预测是社交网络分析领域中非常重要的领域之一，研究者们提出了各种算法来对信息级联的传播进行建模和预测。传统的模型通常难以兼顾预测效果和预测可解释性，而且依赖于大量的标签数据，泛化性能和迁移性能较差。在本文中，我们分别提出了两个创新的信息级联预测模型，从不同的角度来解决传统模型所面临的缺点和局限性。

CasFlow 模型基于一个层级信息级联学习框架，与传统模型单一的建模方式不同，它可以同时对信息级联图的全局和局部结构进行嵌入学习。它引入了语境化下的用户行为学习，对信息传播的时间和结构特征进行了整合，同时考虑了预测的微观角度和宏观角度。CasFlow 模型是第一个考虑了信息传播不确定性的规模预测模型，它使用了变分自编码器和正则化流来对节点级别和级联级别的不确定性进行嵌入学习。它将信息级联的表示当作一种灵活、复杂的后验分布，对节点之间交互行为的概率进行建模。实验结果表明，相比于最先进的基准模型，CasFlow 显著地提升了预测效果，同时还提供了较好的预测行为可解释性。

CCGL 是第一个利用无标签数据、图数据增强、知识蒸馏和迁移学习的信息级联规模预测模型。相比于传统的监督学习基准模型，CCGL 创新地使用了对比自监督预训练来从无标签数据和有标签数据中学习信息级联的通用表示，学习到的表示可以通过知识蒸馏和迁移学习应用到其他类型的数据集和预测任务之中。通过模拟信息在网络中的传播过程，CCGL 设计了针对于信息级联图的数据增强

策略，它可以有效地对数据进行增强和扩充，从而提升模型预测的鲁棒性。CCGL模型不依赖于大量的标签数据，在数据受限的情况下拥有更好的泛化性能和迁移性能。

1.5 本文的结构安排

本文的章节结构安排为五章，主要介绍了研究的问题、研究的背景与意义、国内外已有的研究工作和模型、详细的模型实现描述、详细的实验描述与分析等。本文的具体组织结构安排如下：

第一章 **绪论**：主要介绍本文研究工作的背景与意义、信息级联规模预测的研究历史与现状、本文的主要研究内容、主要的创新与贡献。

第二章 **信息级联规模预测**：针对本文所研究的信息级联规模预测问题，本章对其所涉及的信息级联类型、问题定义、常用评价指标和数据集进行介绍，来帮助读者更好地理解本文中所提出的信息级联预测模型。

第三章 **基于图神经网络的信息级联规模预测**：本章主要介绍 CasFlow 模型的主要内容。首先介绍了与 CasFlow 有关的相关工作，然后定义了该方法中的数学符号，介绍了 CasFlow 模型的框架和具体实现细节，包括基于图小波的信息级联局部结构建模、基于稀疏矩阵分解的信息级联全局结构建模、基于层级变分自编码器和正则化流的不确定性建模、基于双向循环神经网络的信息级联规模预测、以及算法复杂度分析。本章的最后介绍了实验所用到的数据集、实验设置、对比结果及各种实验分析等。

第四章 **基于图对比自监督学习的信息级联规模预测**：本章主要介绍 CCGL 模型的主要内容。首先介绍了与 CCGL 有关的相关工作，然后定义了该方法中的数学符号，介绍了 CCGL 模型的框架和具体实现细节，包括基于图对比自监督学习的信息级联建模、基于微调和知识蒸馏的模型优化、基于双向循环神经网络的信息级联规模预测、以及算法复杂度分析。本章最后介绍了实验所用到的数据集、实验设置、对比结果及各种实验分析等。

第五章 **全文总结与展望**：本章对全文进行了概括和总结，针对当前信息级联规模预测研究工作中存在的问题进行了讨论，并且给出了未来的研究方向。

第二章 信息级联规模预测

信息传播和信息级联预测是社交网络领域中的重要问题，在近些年得到了学界和业界的广泛关注，研究者们提出了各种信息级联预测模型，专注于解决各种特定的现实任务和现实数据。研究信息传播和信息级联预测具有非常重要的经济和社会效益，例如，人们可以通过信息级联预测模型来预测新冠病毒（COVID-19）的感染人数，从而辅助政府和医疗部门做出决策。

由于信息本身天然的多样性，已有的工作从各种不同的角度对其进行建模和预测。信息级联规模预测可以大致从三个角度进行分类：分类预测或者回归预测、在信息发表之后预测或者之前预测、预测的粒度或范围。

本章首先对信息级联的类型做出介绍，然后正式对信息级联规模预测问题、信息级联图和全局图做出定义，最后介绍信息级联规模预测中常见的评价指标和数据集。

2.1 信息级联的类型

在本文中，我们将信息定义为任何可以衡量其规模或流行度的独立存在对象，而信息级联（Information Cascades）由一系列传播的信息序列组成。最著名和最广为研究的信息即为用户生成内容。由于 Web 2.0 服务和移动设备的快速发展，互联网上的信息的产生和阅读方式发生了巨大的改变。传统的信息接受者开始变为信息生产者，他们在各大社交网络平台上发布文字、图片、视频等，这些信息可以通过各种传播机制在互联网上进行快速的传播。信息的产生方法、展示方式、传播途径等都与传统的媒体不同。甚至在传统的媒体领域，例如杂志、报纸、期刊等，也开始在互联网上发布内容。理解信息在网络中的传播方式和机制成为许多现实世界任务中不可或缺的一环，例如广告营销、决策制定、缓存策略等。

信息的传播受到各种内部的和外部的因素影响。例如，在微博和推特平台上，用户可以方便地对其他用户的进行关注、对信息进行评论、“喜欢”，以及对信息进行转发（让自己的关注者看到转发的信息）等。另一方面，用户驱动的分享行为也可以被平台外部的事件所影响，例如爆发事件[62]等。

2.2 信息级联规模预测的问题定义

令 C_k 表示一条信息级联，其发布于一个起始时间，然后在网络中进行传播。在本文的之后部分，为了方便起见，我们默认该信息为一条微博。不过，本文提

出的预测模型可以应用到其他类型的数据上，例如科研论文、新闻文章、在线论坛、多媒体内容等。对于一个用户 u 在时间 t_0 发布的信息 I，其他用户可以对这条信息进行交互，例如：评论、“喜欢”、转发等操作。在本文中，我们将转发行为看作是主要的信息传播方式。对于一个观测时间 t_o，我们假设对于微博 I 在观测时间内共有 M 个用户对其进行转发操作，则一系列转发的微博构成了一个转发信息级联，定义为：

$$C_k(t_o) = \{(v_i, u_i, t_i)\}_{i \in M}, \tag{2-1}$$

其中每一个三元组代表用户 u_i 在时间 $t_i \leq t_o$ 转发了用户 v_i 的微博。

一部分模型 [48,80] 将信息级联预测定义为分类任务，例如，预测信息级联会不会在未来超过现在规模的一倍大小 [12]、预测信息级联的规模会不会超过一个预定的阈值 [81,82]、预测信息级联的规模会增长到哪个区间 [42,83] 等。跟随之前的工作，我们将信息级联规模预测定义为回归任务：

定义一 **信息级联规模预测（Information Cascade Popularity Prediction）**: 给定在时间 t_o 观测到的信息级联 $C_k(t_o)$，信息级联规模预测的目的是预测该信息级联在预测时间 $t_p >> t_o$ 的规模 $P_k(t_p) = |C_k(t_p)|$。以微博为例，预测目标即为转发这条微博的用户数量。

在这个定义的设定下，与观测固定数量的信息级联参与者的方法不同 [12]，我们对信息级联的早期发展观察一段固定的时间 $[t_0, t_o]$。这个设定在现实任务中更为灵活。总的来说，对于 N 个已观测的信息级联（例如 N 个微博）$\{C_k(t_o)\}_{1 \leq k \leq N}$，规模预测可以被定义为回归任务，模型可以通过以下的损失函数来进行训练和优化：

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{k=1}^{N}(\hat{P}_k(t_p) - P_k(t_p))^2, \tag{2-2}$$

$$\hat{P}_k(t_p) = \text{Model}_\Theta(C_k(t_o)), \tag{2-3}$$

其中 $P_k(t_p) = |C_k(t_p)|$ 是我们要预测的最终信息级联 C_k 在预测时间 t_p 的规模，Θ 是模型中的参数。

以上是信息级联规模预测的原始版本。在实际中，模型还可以利用到许多额外的会影响信息级联规模的因素。之前的工作 [84,85] 发现，信息级联文本对信息级联最终的规模有很大的影响。从神经网络中抽取的图像特征也被用来预测信息级联规模 [52]。从社交网络中抽取的特征，例如用户的关注者数量等，也被用来量化用户的影响力 [65,81]。

在本文中，我们主要从两个角度来对信息级联进行建模：信息级联图和用户

社交网络（全局图）。下面正式对它们进行定义：

定义二 **级联图（Cascade Graph）**：给定微博 I 和它对应的转发信息级联 C，信息级联图可以定义为 $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c)$，其中 $\mathcal{V}_c = \{u_i | 1 < i \leq M\}$ 是所有参与信息级联的用户节点，$\mathcal{E}_c \subseteq \mathcal{V}_c \times \mathcal{V}_c$ 是数量为 $M = |C|$ 的边的集合，代表用户在信息级联图中的所有交互关系（在这个例子中是转发关系）。随着时间增长的一个级联图的示意图请见图2-1。

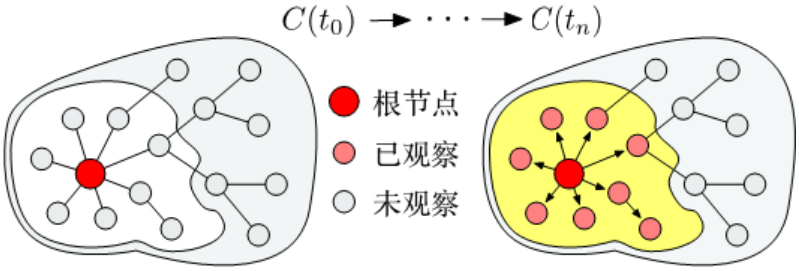


图 2-1 信息级联图

定义三 **全局图（Global Graph）**：全局图中包含着社交网络中所有的节点和边，可以被定义为 $\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$。其中边代表着信息级联中不同的节点关系。一个典型的全局图的例子是微博中用户的关注和被关注网络。

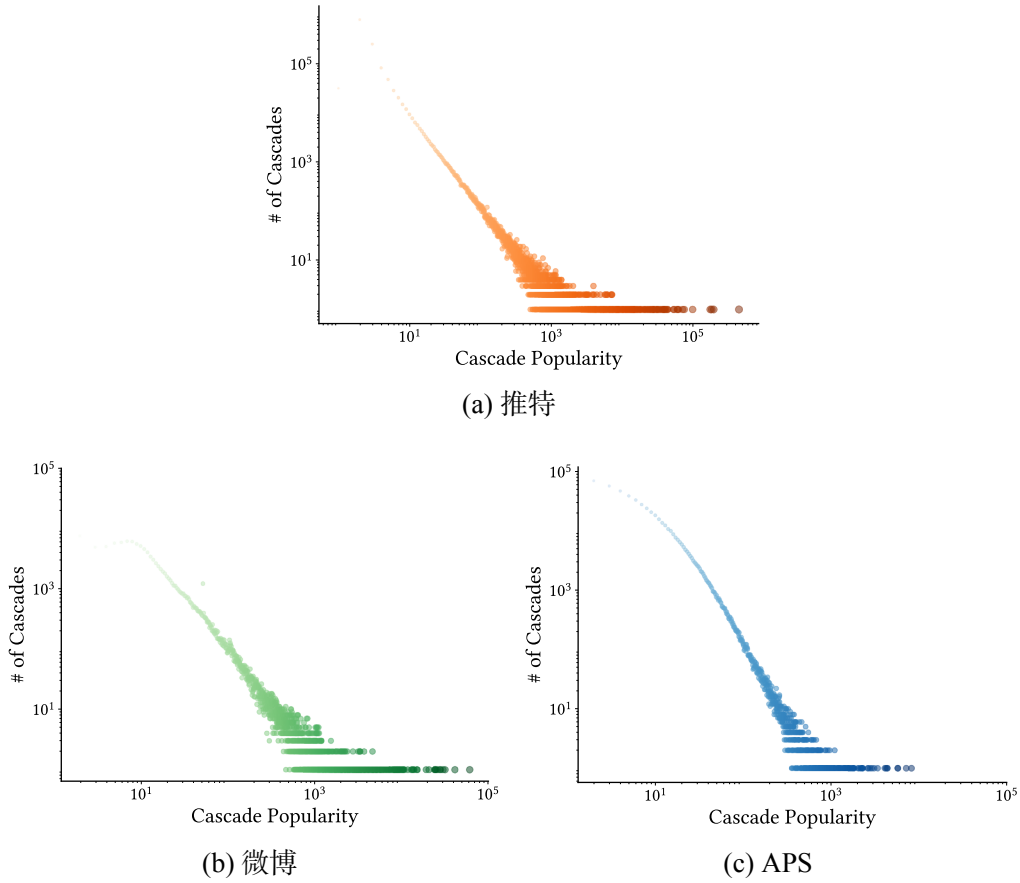
在本文中，信息级联图表示了信息在网络中的局部传播特性，而全局图表示了整个网络中节点之间的关联。以微博为例，用户之间的关注关系、转发关系、历史行为等，都可以反映在全局图的结构之中。之前的工作[7,9]简单地使用了用户的粉丝数（可以看作是节点的度）等特征当作用户的结构特征，它们不能完整地捕捉到用户的影响力、偏好等属性。还有一些其他的工作[46,86]使用了其他类型的结构特征，但是它们也都对信息传播的内在机制做了强假设，或者面临着在特定数据上过拟合的风险，从而导致它们的泛化性能不佳，在迁移到其他应用或数据平台（有着不同的传播机制或传播机制未知）上时效果较差。

2.3 信息级联规模预测中常见的评价指标和数据集

本节介绍信息级联规模预测中常见的评价指标和数据集。

2.3.1 评价指标

对分类任务来说，最常见的评价指标有准确率（Accuracy）、精确度（Precision）、查全率（Recall）和 F 度量（F-measure）等。对于一个预先设置好的阈值，如果其规模超过了这个阈值，信息级联可以被分为爆发级联。准确率评价指标的



(a) 推特

(b) 微博

(c) APS

图 2-2 信息级联规模分布。依次分别为推特数据集、微博数据集和 APS 数据集。X 轴和 Y 轴分别进行了对数化操作。我们对数据进行幂律分布拟合，三个数据集的指数 α 分别为 1.916、2.044 和 2.788，其中最小规模 $P_{\min}$ 分别大于或等于 34、45 和 58.

一大缺陷是在数据的类（Class）极度不平衡的时候（规模分布不均）无法有效地评价模型的预测效果。以推特数据集 [87] 为例，其中超过 92.8% 的推文的规模小于 10，只有 0.114% 的推文的规模超过了 1,000。这些推文的规模服从于厚尾分布（Heavy-tailed distribution），如图2-2所示。

对于回归任务来说，最常见的评价指标是均方误差（Mean Square Error, MSE）和其变种。规模或流行度通常用对数尺度（Logarithmic scale）来计算，从而避免损失函数或评价指标被离群值所影响，以及保证计算过程中的数值稳定，例如均方对数误差（Mean Square Logarithmic Error）。测定系数（Coefficient of determination）、相关系数（Coefficient of correlation）及它们的变种等评价指标也经常被使用。

2.3.2 数据集

信息级联所蕴含的数据范围广阔，它们可以包括新闻文章、科研论文、社交媒体上的图片、音乐、视频等。这种内容多样性使得规模预测模型的设计变得复杂。常见的信息级联规模预测数据集包括推特（Twitter）、微博（Weibo）、脸书（Facebook）、掘客（Digg）、相薄（Flickr）、YouTube、APS、DBLP 等。

2.3.3 本章小结

本章具体介绍了信息级联规模预测问题的类型、问题定义、常见的评价指标和数据集等，为读者理解第三章和第四章中提出的两个预测模型提供了必要的预备知识。

第三章 基于图神经网络的信息级联规模预测

本章主要介绍基于图神经网络的信息级联规模预测模型 CasFlow 的实现细节以及相关实验。CasFlow 模型对信息在网络中的局部传播和全局传播进行基于概率的不确定性建模，同时考虑了信息级联的时序特征和结构特征，它对语境化下的用户行为进行层级学习，捕获到了节点级别和级联级别的传播不确定性，显著地提升了预测效果。本章首先讨论与 CasFlow 模型有关的相关工作，并解释了 CasFlow 模型与它们的不同，以及 CasFlow 模型如何解决它们所面临的缺点和局限性。然后我们对 CasFlow 模型的细节进行介绍，主要包括（1）基于图小波和稀疏矩阵分解的信息级联的结构学习；（2）基于双向门控循环单元的信息传播的时序特征学习；（3）基于变分自编码器和正则化流的层级传播不确定性建模；以及（4）信息级联规模预测器。然后我们对 CasFlow 模型的计算复杂度进行了分析。随后我们介绍了本文对 CasFlow 模型所做的实验内容，包括数据集处理、基准模型、实验参数设置、评价指标、实验环境、实验结果等。我们还对 CasFlow 模型做了大量的消融实验，并描述了模型预测的可解释性。表3-1中介绍了本章中所常用的数学符号。

3.1 相关工作介绍

基于特征工程的模型从与信息级联有关的各种影响信息级联规模的因素中构建特征，例如与内容有关的特征（标签和提醒的数量）、与用户有关的特征（用户主页、属性、历史行为等）、信息级联中包含的文本和图像、信息级联的时序和结构特征等。在选择和构建好特征集合后，特征工程模型使用各种机器学习技术来对信息级联的规模作出预测。在许多工作中[37,88]，作者们发现时序特征和用户特征对信息级联规模预测更为有效。基于特征工程的模型的泛化性能较差，因为选取特征的过程依赖于专家知识、特定的数据来源和特定的应用，而当前并没有一个系统的方法或指南来引导特征的选取和设计。CasFlow 模型不依赖于特征工程，是一个端到端的由数据驱动的基于图神经网络的模型，它的泛化性能较高，可以自动地从数据中学习到对准确预测有用的信息表示，避免了对特征的复杂特殊设计和人工偏向。

基于统计和概率的生成模型对信息级联中的时序特性和模式进行学习。它们对事件的到达过程（Arrival process）用一种生成式的方式来建模。一般来说，信息级联被看作是时间序列数据，模型通过对观测时间内的事件发生概率进行参数

表 3-1 本章中所用到的数学符号及其描述

符号	描述
$\boldsymbol{A}, \boldsymbol{D}$	邻接矩阵和对角度矩阵。
$C_k(t)$	在时间 t 观测到的信息级联。
d	嵌入的维度.
$E_c(u_i)$	级联图中节点 u_i 的嵌入表示。
$E_g(u_i)$	全局图中节点 u_i 的嵌入表示。
$\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c)$	由一系列节点和边组成的级联图。
$\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$	由一系列节点和边组成的全局图。
$\boldsymbol{h}_1, \boldsymbol{h}_2$	双层双向门空循环单元中的向量表示。
I	信息（例如微博或论文）。
K	正则化流中变换的数量。
M	信息级联中用户的数量。
N	信息级联的数量。
$P_k(t)$	时间 t 时信息级联的规模。
$\mathcal{R} = \{\mathcal{R}_i\}_{i \in \lvert\mathcal{V}_c\rvert}$	变分自编码器的输入序列。
t_o, t_p	观测时间和预测时间。
u_i	信息级联中的用户。
$\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3$	分别来自节点级别的变分自编码器、级联级别的变分自编码器、以及正则化流的隐藏表示。

最大化估计。各种点过程（Point process，例如柏松过程和霍克斯自激励过程）和统计模型（例如 Cox、威布尔分布、生存分析、传染模型等）被用来学习信息传播中隐含的传播机制。生成模型的预测可解释性强，但是并不能完全利用信息级联中隐式的动态变化。在最近的一篇信息级联预测综述[89]中，作者们发现简单的柏松过程难以捕获传播的模式，而威布尔和霍克斯自激励模型倾向于过高估计信息级联的规模，其背后的原因很可能在与它们所采用的自激励机制。CasFlow 模型对信息传播过程中的时序特征和结构特征进行有效的学习，不依赖于特定的、假设的传播机制。

最近在许多领域取得成功的深层神经网络启发了许多基于深度学习的信息级联预测模型，它们显著地提升了已有模型的预测效果。早期的 DeepCas 是一个基于信息级联结构的规模预测模型[65]，它利用随机游走方法来对信息级联图的结构进行端到端的表示学习。之后的 DeepHawkes 模型[21]把信息级联图分解为多个传播路径，每个路径表示用户之间基于时间的信息传播方向。此外还有很多基于深度学习的预测模型，例如 DTCN[73]、UHAN[90]、Topo-LSTM[91]、FOREST[92] 和 DFTC[83] 等，它们从结构和序列建模的角度来学习信息传播的特性。循环神经网

络和注意力机制被用来对信息规模增长进行建模。与上述专注于简单图表示学习的模型不同，CasFlow 同时学习了局部的和全局的信息级联结构特性。

循环级联卷积模型（Recurrent Cascade Convolution Model, CasCN）[76] 使用一个动态的图卷积网络（Graph Convolutional Network, GCN) 来对信息级联的结构进行表示学习，它还考虑到了信息级联传播的方向性和时间衰减因素。DMT-LIC 模型 [93] 从多任务学习（Multi-task learning）的角度来学习信息级联的表示，特别地，它从宏观的角度来预测信息级联的规模，同时从微观的角度来预测用户的分享行为。Coupled-GNN 模型 [77] 利用了耦合的两个图神经网络来学习信息传播中的级联现象。这些模型依赖于确定性的推理过程，限制了它们对信息级联中变化的建模，而 CasFlow 模型则考虑了信息传播中的不确定性。

3.2 CasFlow 模型总体框架

在本节中，我们介绍 CasFlow 模型的总体框架。模型的框架图见图3-1，实现代码见https://github.com/Xovee/casflow。它主要包含以下四个部分：

（A） **结构学习**：CasFlow 主要对信息级联图中语境化的结构模式和用户在社交网络中的隐含关系进行捕捉和建模。它利用了图信号处理（Graph signal processing [94,95]）中的技术来学习信息级联的结构表示：基于小波图谱（Spectral graph wavelets）的局部结构建模和基于稀疏矩阵分解（Sparse matrix factorization）的用户全局结构建模。

（B） **时间传播建模**：CasFlow 使用了双向循环神经网络来对信息传播中的时序依赖进行建模。

（C） **传播不确定性建模**：CasFlow 使用了变分自编码器（Variational Auto-Encoder, VAE）对信息传播和信息增长中的变化和不确定性进行了建模，并且用正则化流（Normalizing Flows, NFs）来对隐藏变量的预估后验分布（Posterior distribution）进行了一系列复杂和灵活的转化。

（D） **预测器**：CasFlow 结合了循环神经网络和变分推断（Variational inference）来学习信息级联的高阶表示，最后使用多层感知机（Multi-Layer Perceptrons, MLPs）来对信息级联的最终规模作出预测。

3.3 基于图神经网络的信息级联结构学习

3.3.1 语境化下的信息级联图学习

为了捕获信息中的局部结构特征，我们使用了一种图嵌入技巧，即图谱小波（Spectral graph wavelet）[96] 来学习每个用户在信息级联图中的结构表示。其他类

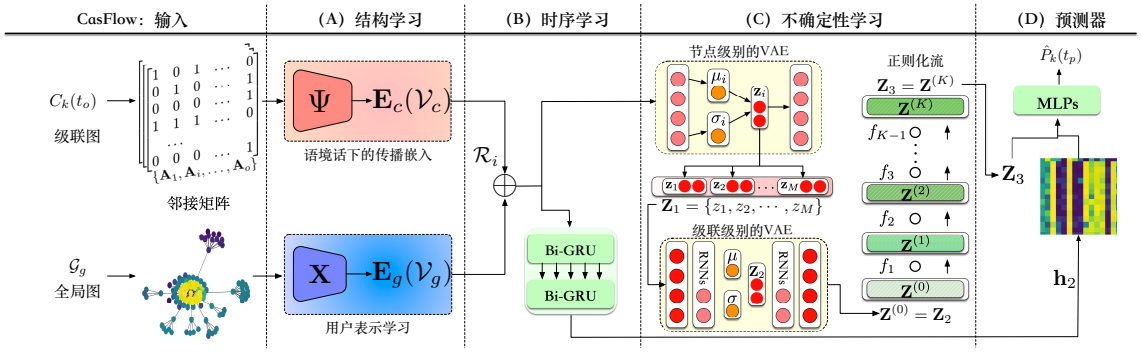


图 3-1 CasFlow 模型的总体架构。t_0 是信息的发布时间，t_o 是信息的观测时间，t_p 是信息的预测时间。模型的输入包括信息级联图和全局图。（A）：基于语境化的信息级联传播局部结构建模和全局图中的用户结构表示学习；（B）：基于双层双向门空循环单元的时序建模；（C）：基于层级变分自编码器和正则化流的传播不确定性建模；（D）：基于多层感知机的信息级联规模预测器。

型的图表示学习也可以在这里使用，例如 DeepWalk [66]、node2vec [72] 等，取决于具体的结构学习目标。

给定一个微博 $C \in \{C_1, C_2, \cdots C_N\}$ 和它在 t_o 时刻观测到的信息级联图 $\mathcal{G}_c(t_o)$，它的加权邻接矩阵（Weighted adjacency matrix）可以被定义为 $\boldsymbol{A}_c$。对角度矩阵（Diagonal degree matrix）$\boldsymbol{D}_c$ 中每一个对角元素等于所有连接到当前节点 u_i 的边的权重之和。我们随之定义未标准化的图拉普拉斯算子（Unnormalized graph Laplacian）：

$$\boldsymbol{L}_c = \boldsymbol{D}_c - \boldsymbol{A}_c = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T, \tag{3-1}$$

其中 $\boldsymbol{U}$ 是特征值分解，$\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_0, \ldots, \lambda_{M-1})$ 是满足于 $\lambda_0 < \lambda_1 \leq \cdots \leq \lambda_{M-1}$ 条件的对角矩阵的特征值。我们随后可以将每个节点 $u_i \in \mathcal{V}_c(t_o)$ 的图谱小波 $\Psi_{u,s}$ 计算为：

$$\Psi_{u,s} = \boldsymbol{U}\,\mathrm{Diag}(g_s(\lambda_0), \ldots, g_s(\lambda_{M-1}))\boldsymbol{U}^T\delta_u, \tag{3-2}$$

其中 δ_u 是节点 u 的 one-hot 编码向量，滤波核 g_s 是定义在 $\mathbb{R}^+$ 上的连续函数。这里我们使用热核函数（Heat kernel function）$g_s(\lambda) = e^{-\lambda s}$，其中 s 是一个在频谱 $(\lambda_l)_{l=0,\ldots,M-1}$ 上定义的尺度参数。

图拉普拉斯算子的特征值和特征向量的含义与信号的频率拥有相似的概念，也就是说，与大的特征值相关联的特征向量在图中变化的更快，因此，这些特征向量在这些位置上倾向于拥有不同的数值 [95]。相反，与小的特征值相关联的特征

向量拥有在边之间变化较慢的信号，使得相互权值较大的邻居节点倾向于拥有相似的数值。而我们使用的热核函数 g_s 直接定义在图谱域上，它内部的低通滤波器使得数值从高到低的变化拥有一个相对平滑的改变。

对节点的嵌入最基本的观点是让小波的系数直接与图的拓扑属性（Graph topological properties）相关，因此，节点的嵌入会包含足够的信息来对结构上相似的节点进行恢复[96]。对于一个给定的节点 u_i，我们将它的小波系数看作是一种概率分布，然后使用经验特征函数（empirical characteristic functions）[97] 来表示这个分布。对于一个为标量的随机变量 X，它的特征函数定义为 $\varphi_X(p) = \mathbb{E}[e^{ipX}], p \in \mathbb{R}$。特别地，对于一个给定的节点 u_i 和一个尺度参数 s，经验特征函数由下面的公式正式定义：

$$\varphi_{u,s}(p) = \frac{1}{M}\sum_{m=1}^{M} e^{ip\Psi_{m,u,s}}, \Psi_{m,u,s} = \sum_{l=0}^{M-1} g_s(\lambda_l)U_{ml}U_{ul}, \tag{3-3}$$

其中 $\Psi_{m,u,s}$ 是 $\Psi_{u,s}$ 的第 m 个小波系数。然后信息级联图中节点 u_i 的嵌入可以通过串联实数部分和虚数部分来得到：

$$E_c(u_i) = [\mathrm{Re}\ \varphi_{u,s}(p), \mathrm{Im}\ \varphi_{u,s}(p)]_{p_1,p_2,\ldots,p_d}. \tag{3-4}$$

节点嵌入 $E_c(u_i)$ 的维度为 $d_c = 2d$，嵌入的第一个元素被设置为节点边的权值，该权值通过下面的公式来定义和正则化：

$$W_u = (t_j - t_o)/t_o \in [0,1], 0 < t_o \leq t_j. \tag{3-5}$$

除了图中节点的表示，使用小波来学习结构信息与网络中信息的传播过程类似，这使得我们可以对基于上下文的用户行为进行建模，也就是说，我们关注的是独立节点嵌入，而不是对整个信息级联图的嵌入或者是聚焦于特定的任务。在另一方面，我们转向学习全局图中的用户表示，这种表示表达了用户之间的连通性，并且隐含了用户的历史行为特征。

3.3.2 在大规模全局图中进行可扩展的表示学习

与信息级联图 $\mathcal{G}_c$ 不同，全局图 $\mathcal{G}_g$ 通常包含多达百万计的节点和边，这使得对其进行表示学习变得非常困难。已有的图学习模型[72,77] 较难直接应用到实际的信息级联预测问题中。在本节中，我们使用稀疏矩阵分解（Sparse matrix factorization）[98] 来高效地、可扩展地处理和建模大规模的全局图。

给定全局图 $\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$，其可以定义为一个社交网络（例如，关注和被关注网络），或者定义为一个用户之间的交互网络（例如，“喜欢”、提醒、转发网络），

或者两者的组合。$\boldsymbol{A}_g$ 是其加权邻接矩阵，$\boldsymbol{D}_g$ 是其对角度矩阵。特别地，为了避免大规模矩阵的因式分解的不可计算性，我们使用了稀疏随机化截断奇异值分解（Sparse Randomized Truncated Singular Value Decomposition, TSVD）来学习基于分布相似度度量（Distributional similarity-based）的节点嵌入，这种方法兼顾了计算效率和学习效果[99,100]。特别地，根据 [98]，关系矩阵（Proximity matrix）$\boldsymbol{X}$ 可以被定义为：

$$\boldsymbol{X}_{i,j} = \begin{cases} \ln p_{i,j} - \ln(\tau Q_{\mathcal{E}_g,j}), & (u_i, u_j) \in \mathcal{E}_g, \\ 0, & (u_i, u_j) \notin \mathcal{E}_g, \end{cases} \tag{3-6}$$

其中 τ 是负样本比率，$p_{i,j}$ 是 $\mathcal{E}_g$ 中用户对 (u_i, u_j) 的权值，以及 $Q_{\mathcal{E}_g,j}$ 是节点 u_j 的负样本。然后优化的目标变为对 $\boldsymbol{X}$ 的矩阵分解近似：

$$\boldsymbol{X} \approx \tilde{\boldsymbol{X}} = \boldsymbol{U}_{d_g} \boldsymbol{\Sigma}_{d_g} \boldsymbol{V}_{d_g}^T, \tag{3-7}$$

其中 $\boldsymbol{U}_{d_g}, \boldsymbol{V}_{d_g} \in \mathbb{R}^{|\mathcal{V}_g| \times d_g}$ 是正交矩阵（Orthonormal matrices），$\boldsymbol{\Sigma}_{d_g}$ 是一个矩形对角矩阵，其中有 d_g 个选择的非负奇异值。因为 $d_g \ll |\mathcal{V}_g|$，计算的复杂度极大地减少了。

为了进一步加速大规模图的计算效率，我们使用了随机化的截断奇异值分解来近似矩阵 $\boldsymbol{X}$，计算分为两步：（1）首先我们尝试去寻找满足拥有 d_g 个正交列的 $\boldsymbol{R}$，然后让 $\boldsymbol{X} \approx \boldsymbol{R}\boldsymbol{R}^T\boldsymbol{X}$（为了方便起见，我们省略了下标）；（2）假设我们找到了满足条件的 $\boldsymbol{R}$，然后使 $\boldsymbol{B} = \boldsymbol{R}^T\boldsymbol{X} \in \mathbb{R}^{d_g \times |\mathcal{V}_g|}$，这是一个相对较小的矩阵，可以使用标准的奇异值分解算法来有效率地进行计算，然后我们有：

$$\boldsymbol{B} = \boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{V}^T, \tag{3-8}$$

其中 $\boldsymbol{S}, \boldsymbol{V}$ 是正交的，以及 $\boldsymbol{\Sigma}$ 是对角的，最后令 $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{S}$ 来获取最终的近似矩阵：

$$\boldsymbol{X} \approx \boldsymbol{R}\boldsymbol{R}^T\boldsymbol{X} = \boldsymbol{R}(\boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{V}^T). \tag{3-9}$$

为了快速地找到矩阵 $\boldsymbol{R}$，我们首先生成高斯随机矩阵 $\boldsymbol{\Omega} \in \mathbb{R}^{|\mathcal{V}_g| \times d_g}$，计算 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Omega}$，然后对 $\boldsymbol{Y}$ 进行 QR 分解（QR decomposition）。最后，$\mathcal{V}_g$ 中节点的嵌入可以通过以下公式获得：

$$\boldsymbol{E}_g(\mathcal{V}_g) = \{E_g(u_i)\}_{u_i \in \mathcal{V}_g} = \boldsymbol{R}_{d_g} \boldsymbol{S}_{d_g} \boldsymbol{\Sigma}_{d_g}^{1/2}. \tag{3-10}$$

与信息级联图中的节点嵌入 $\boldsymbol{E}_c(\mathcal{V}_c)$ 相比（具体请见节3.3.1），全局图中的节点嵌入 $\boldsymbol{E}_g(\mathcal{V}_g)$ 表达了全局图中截然不同的信息传播概念。对于信息级联图来说，

不管对于那些有影响力的节点来说，还是对于那些连接不同社区的枢纽节点来说，还是对于不起眼的叶子节点来说，拥有相似结构位置的节点将会拥有相似的节点嵌入，就算它们在图中的距离非常远。这一位置特性由图小波的传播模式来捕获。对于全局图来说，模型学习到的低维度的连续嵌入保留了节点的在全局图中的邻进性（Proximity），因此，拥有相似偏好和行为的节点将会拥有相似的空间嵌入。

3.4 时序传播建模

在上一节中，我们使用了图小波和稀疏矩阵分解来生成编码了用户在信息级联图 $\mathcal{G}_c$ 和全局 $\mathcal{G}_g$ 图中的结构信息嵌入。特别地，它们由如下特征：（1）信息级联图中结构上等同的节点将会拥有相似的嵌入（参考 [96]），例如，枢纽节点比叶子节点拥有更强大的传播能力；（2）在全局图中临近的节点将会拥有相似的嵌入，也就是说，临近的节点们拥有相似的偏好来传播特定的信息。

除了信息级联中蕴含的结构信息，时序信息被认为是信息级联规模预测问题中最为重要的一类特征之一，对信息级联最终的规模有着关键的影响。为了捕获信息级联的时序特性，我们使用了双向门控循环单元（Bi-directional Gated Recurrent Units, Bi-GRU）来对信息中的级联效应进行建模。循环神经网络广泛地应用到时序数据的建模之中，例如在 [91] 和 [101] 中，循环网神经网络被用来对信息传播中的时序特征进行建模。

对于一个给定的信息级联 C，我们有 $|\mathcal{V}_c|$ 个节点嵌入 $\boldsymbol{E}_c(\mathcal{V}_c) = \{E_c(u_i)\}_{i\in|\mathcal{V}_c|}$，这些嵌入是我们使用图小波技术在信息级联图上通过预训练而得到的。对于 $\mathcal{V}_c$ 中的每一个节点 u_i，如果 u_i 也在全局图中，也就是说，$u_i \in \mathcal{V}_g$，那么我们已经使用稀疏矩阵分解获取了其在全局图中的嵌入 $E_g(u_i)$。如果 u_i 不在全局图中，我们初始化一个嵌入 $E_g(u_i) = 0 \in \mathbb{R}^{d_g}$ 作为冷启动。随后，节点的嵌入按照时间顺序送到双层双向门控循环单元中，来生成上下文依赖的高阶表示。对于每个输入 $E_c(u_i)$ 和 $E_g(u_i)$，门控循环单元使用门控单元来计算隐藏状态（Hidden state）的更新值。将门控循环单元的前向输出 $\overrightarrow{\text{GRU}}$ 和后向输出 $\overleftarrow{\text{GRU}}$ 串联起来，信息级联最终的表示 $\boldsymbol{h}_2$ 可以由下面的公式计算：

$$\boldsymbol{E} = \text{Concat}(E_c(u_i), E_g(u_i)), \overrightarrow{\boldsymbol{h}}_1 = \overrightarrow{\text{GRU}}(\boldsymbol{E}), \tag{3-11}$$

$$\overleftarrow{\boldsymbol{h}}_1 = \overleftarrow{\text{GRU}}(\boldsymbol{E}), \boldsymbol{h}_1 = \text{Concat}(\overrightarrow{\boldsymbol{h}}_1, \overleftarrow{\boldsymbol{h}}_1), \overrightarrow{\boldsymbol{h}}_2 = \overrightarrow{\text{GRU}}(\boldsymbol{h}_1), \tag{3-12}$$

$$\overleftarrow{\boldsymbol{h}}_2 = \overleftarrow{\text{GRU}}(\boldsymbol{h}_1), \boldsymbol{h}_2 = \text{Concat}(\overrightarrow{\boldsymbol{h}}_2, \overleftarrow{\boldsymbol{h}}_2), \tag{3-13}$$

其中 $\overrightarrow{\boldsymbol{h}}_1$ 和 $\overleftarrow{\boldsymbol{h}}_1$ 是门控循环单元的完整序列输出，$\overrightarrow{\boldsymbol{h}}_2$ 和 $\overleftarrow{\boldsymbol{h}}_2$ 是最后一层的输出向量。到现在为止，如许多之前的工作一样 [65,76]，信息级联表示 $\boldsymbol{h}_2$ 可以被用来对信

息级联的规模做出预测。我们把当前的这个模型成为 CasFlow-RNN。

但是，只使用循环神经网络的最后一层隐藏状态对于信息级联预测来说有一定的缺陷。这是由于循环神经网络中扁平的序列生成过程造成的，每一个节点的嵌入依赖于前一时间的节点嵌入。这个问题在于模型被强迫地用确定性的、逐步生成的方式去生成所有的高阶信息。这一设定对探索信息级联中不确定的依赖具有很大的局限性。另外，受限于循环神经网络自身的局限性，这些模型不能处理长期依赖，当信息级联的长度很长时，它们的预测效果可能会显著下降。

3.5 信息传播不确定性建模

在本节中，我们设计了一个深度生成模型来学习信息传播中的不确定性。为了达到这一点，我们采用了变分自编码器（Variational Auto-Encoder, VAE）[102] 来对传播不确定性进行建模。变分自编码器是一种生成式的模型，它包括一个编码器和解码器，提供了一种通用的学习隐藏表示的框架。它构建了一种数据联合概率分布（Joint probability distribution），然后对隐藏随机变量的后验分布进行学习。学习到的表示可以被用来生成数据，或者被用作其他任务，例如分类 [103]、节点表示 [104]、预测 [105] 和推荐 [106] 等。作为一个基于概率的贝叶斯模型，变分自编码器提供了一种基于统计和数学的角度来解决信息传播中的随机性和不确定性。这启发了我们使用这类贝叶斯框架（Bayesian framework）来对信息级联中的不确定性进行建模。

3.5.1 节点级别（低阶）的不确定性建模

一个信息级联 C 是由一个不断增长的参与者序列组成的，每一个参与者都关联着一个学习到的表示，该表示代表着信息传播的一个特定的阶段。在节3.3.1和节3.3.2中，对于信息级联图和全局图中的每一个节点，我们分别使用图小波和稀疏矩阵分解来对节点学习其嵌入表示 $E_c(u_i)$ 和 $E_g(u_i)$。在更一般的意义上，任何其他类型的图表示学习方法都可以用来增强模型的学习能力，例如，文本和图像的嵌入。在不引起歧义的情况下，我们使用 $\mathcal{R}_i, (i \in |\mathcal{V}_c|)$ 来表示信息级联 C 中每一个参与者，也就是说，$\mathcal{R}_i = \text{Concat}(E_c(u_i), E_g(u_i))$。

令 $\mathbf{Enc}(\cdot)$ 作为输入的编码器，$\mathbf{Dec}(\cdot)$ 作为重构输入的解码器，基于神经网络的深度变分自编码器可以被定义为：

$$z_i = \mathbf{Enc}(\mathcal{R}_i), \overline{\mathcal{R}}_i = \mathbf{Dec}(z_i), \text{ for } i = 1, 2, \ldots, M, \tag{3-14}$$

$$\mu_i = \text{NN}(\mathcal{R}_i), \log \sigma_i^2 = \text{NN}(\mathcal{R}_i), z_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \tag{3-15}$$

其中 $\overline{\mathcal{R}}_i$ 是重构的输入，$z_i \in \mathbb{R}^{d_z}$ 是隐藏向量。变分自编码器接受高维度的数据作为输入，然后生成压缩后的隐藏表示，该表示从一个条件先验分布（Conditional prior distribution）采样而来，该分布的标准差为 μ，方差为 $\log\sigma^2$。然后从该隐藏表示中重构原始的输入。

为了从信息级联数据中学习到基于概率的有效表示，从而捕获信息级联在网络中传播的变化和不确定性，变分自编码器从编码器的输出向量中采样 μ 和 $\log\sigma^2$，然后利用重参数化技巧（Reparameterization trick）来从高斯分布中采样隐藏向量[102]：

$$z_i = \mu_i + \sigma_i\varepsilon, \varepsilon \sim \mathcal{N}(0,1). \tag{3-16}$$

以及，对于信息级联中每个参与者的表示，它对于 $\mathcal{R}_i$ 的边缘对数似然（Marginal log-likelihood）是：

$$\log p_\theta(\mathcal{R}_i) = \log \int p_\theta(\mathcal{R}_i|z_i)p(z_i)\mathrm{d}z_i, \tag{3-17}$$

但是，这个对数似然在隐藏表示中拥有很高维度的时候不能被快速地进行计算。除了难以计算 $\log p_\theta(\mathcal{R}_i)$，通过观察一个参数化的先验 $q_\varphi(z_i|\mathcal{R}_i)$ 来最大化证据下界（Maximizing the Evidence Lower Bound, ELBO）等同于或近似于真实的后验分布 $p_\theta(z_i|\mathcal{R}_i)$：

$$\begin{aligned}
\log p_\theta(\mathcal{R}_i) &= \log \int p_\theta(\mathcal{R}_i|z_i)p(z_i)\mathrm{d}z_i && (3\text{-}18)\\
&= \mathbb{E}_{q_\varphi(z_i|\mathcal{R}_i)} \log\left[\frac{p_\theta(\mathcal{R}_i, z_i)}{q_\varphi(z_i|\mathcal{R}_i)}\right] + \mathbb{D}_{\mathrm{KL}}(q_\varphi(z_i|\mathcal{R}_i)||p_\theta(z_i|\mathcal{R}_i)) && (3\text{-}19)\\
&\geq \mathbb{E}_{q_\varphi(z_i|\mathcal{R}_i)}\left[\log p_\theta(\mathcal{R}_i, z_i) - \log q_\varphi(z_i|\mathcal{R}_i)\right] && (3\text{-}20)\\
&\triangleq \mathrm{ELBO}(\mathcal{R}_i), && (3\text{-}21)
\end{aligned}$$

其中 $q_\varphi(z_i|\mathcal{R}_i)$（也被称为拥有参数 φ 的编码器）是一个真实后验分布 $p_\theta(z_i|\mathcal{R}_i)$ 的近似，该近似被用来生成隐藏向量 z_i，$\mathbb{D}_{\mathrm{KL}}(\cdot)$ 是 Kullback-Leibler 散度（KL divergence），具体定义为：

$$\begin{aligned}
\mathbb{D}_{\mathrm{KL}}(q_\varphi(z_i|\mathcal{R}_i)||p_\theta(z_i|\mathcal{R}_i)) &= \sum_{z_i} q_\varphi(z_i|\mathcal{R}_i)\log\frac{q_\varphi(z_i|\mathcal{R}_i)}{p_\theta(z_i|\mathcal{R}_i)} && (3\text{-}22)\\
&= \mathbb{E}\left[\log\frac{q_\varphi(z_i|\mathcal{R}_i)}{p_\theta(z_i|\mathcal{R}_i)}\right] && (3\text{-}23)\\
&= \mathbb{E}\left[\log q_\varphi(z_i|\mathcal{R}_i) - \log p_\theta(z_i|\mathcal{R}_i)\right]. && (3\text{-}24)
\end{aligned}$$

因为现在的目标是去最小化 $q_\varphi(z_i|\mathcal{R}_i)$ 和 $p_\theta(z_i|\mathcal{R}_i)$ 之间的 KL 散度，我们可以转而

最大化 $\log p_\theta(\mathcal{R}_i, \boldsymbol{z}_i)$ 的证据下界，参数 θ 和 φ 均可以由不同的非线性函数来进行计算，例如神经网络等。

通过最小化输入 $\mathcal{R}_i$ 和输出 $\overline{\mathcal{R}}_i$ 之间的重构损失，从信息级联 C 中所有的参与者里学习到的隐藏表示 $\boldsymbol{Z}_1 = \{\boldsymbol{z}_i\}_{i\in|\mathcal{V}_c|}$ 捕捉了数据中的分布，并且可以用来生成数据或者用来提升特定任务的效果[103,106]。现在，我们可以结合生成的 $\boldsymbol{Z}_1$ 和双层双向门控循环单元来预测信息级联的最终规模，我们将这个模型称作 CasFlow*，它可以被看作是一种基于时序建模的节点级别（低阶）的变分推断模型。如我们将在实验一节所示，这个模型提升了预测的效果，但是它的缺点在于只能捕捉独立的节点级别的不确定性，忽略了信息级联中动态变化的不确定性。另外，低阶的变分推断丢弃了信息级联中参与者之间的序列依赖。

3.5.2 级联级别（高阶）的变分推断

为了克服上述 CasFlow* 模型的浅层生成问题，如图3-1所示，我们将循环神经网络结合到序列变分自编码器中。这一个高阶的级联级别的变分自编码器接受 M 个由低阶变分自编码器生成的序列隐藏变量 $\boldsymbol{Z}_1 = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M\}$ 作为输入，每一个 $\boldsymbol{z}_i$ 对应着级联 C 中的一个参与者，以此来最小化重构损失。我们由此可以获得级联级别的隐藏表示 $\boldsymbol{Z}_2$，该表示同时捕获了时间和序列关系，因此表达了信息级联不断变化地轨迹中信息传播的因果关系和依赖。

然后，让 $\mathbf{Enc}(\cdot)$ 表示基于循环神经网络的编码器输入，以及让 $\mathbf{Dec}(\cdot)$ 表示基于循环神经网络的解码器来对输入进行重构，则基于循环神经网络的变分自编码器可以被正式定义为：

$$\boldsymbol{Z}_2 = \mathbf{Enc}(\boldsymbol{Z}_1), \overline{\boldsymbol{Z}}_1 = \mathbf{Dec}(\boldsymbol{Z}_2), \tag{3-25}$$

$$\begin{aligned} \boldsymbol{Z}_1 = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M\}, \overline{\boldsymbol{Z}}_1 = \{\overline{\boldsymbol{z}}_1, \overline{\boldsymbol{z}}_2, \ldots, \overline{\boldsymbol{z}}_M\}, \\ \mu = \mathrm{NN}(\mathrm{RNN}(\boldsymbol{Z}_1)), \log\sigma^2 = \mathrm{NN}(\mathrm{RNN}(\boldsymbol{Z}_1)), \end{aligned} \tag{3-26}$$

$$\boldsymbol{Z}_2 \sim \mathcal{N}(\mu, \sigma^2), \overline{\boldsymbol{z}}_i = \mathrm{NN}(\mathrm{RNN}(\boldsymbol{Z}_2)), \text{ for } i = 1, 2, \ldots, M, \tag{3-27}$$

其中 $\overline{\boldsymbol{Z}}_1$ 是重构的输入，M 是序列的长度，$\boldsymbol{Z}_2 \in \mathbb{R}^{d_{\boldsymbol{Z}_2}}$ 是学习到的压缩的隐藏向量。更准确地说，使 $\mathcal{R} = \{\mathcal{R}_i\}_{i\in|\mathcal{V}_c|}$ 表示输入序列，则信息级联 C 的联合概率可以被定义为：

$$p_\theta(\mathcal{R}, \boldsymbol{Z}_1, \boldsymbol{Z}_2) = p_\theta(\boldsymbol{Z}_2|\boldsymbol{Z}_1) p_\theta(\mathcal{R}|\boldsymbol{Z}_1, \boldsymbol{Z}_2), \tag{3-28}$$

其中隐藏向量 $\boldsymbol{Z}_1$ 是中心各向同性多元高斯分布（Centered isotropic multivariate Gaussian distributions）。则条件分布 $p(\boldsymbol{Z}_2|\boldsymbol{Z}_1)$ 的参数由一个基于循环神经网络的编

码器来确定，$p(\mathcal{R}|\boldsymbol{Z}_1,\boldsymbol{Z}_2)$ 可以被考虑作从隐藏因子中重构的信息级联表示，这个过程可以正式定义为：

$$p_\theta(\boldsymbol{Z}_2|\boldsymbol{Z}_1)=\sum_{m=1}^{M}\mathcal{N}(\boldsymbol{Z}_2|f_\vartheta^{\mu}(\boldsymbol{Z}_1)),\mathrm{Diag}(f_\vartheta^{\sigma^2}(\boldsymbol{Z}_1))), \tag{3-29}$$

$$p_\theta(\mathcal{R}|\boldsymbol{Z}_1,\boldsymbol{Z}_2)=\mathcal{N}(\mathcal{R}|f_\phi^{\mu}(\boldsymbol{Z}_1,\boldsymbol{Z}_2),\mathrm{Diag}(f_\phi^{\sigma^2}(\boldsymbol{Z}_1,\boldsymbol{Z}_2))), \tag{3-30}$$

其中已观测到的级联 $\mathcal{R}$ 的条件分布是具有对角协方差矩阵的多元高斯函数（Multivariate Gaussian with a diagonal covariance matrix），其均值和对角协方差由神经网络 f_*^{μ} 和 $f_*^{\sigma^2}$ 来确定，ϑ 和 ϕ 是神经网络的参数。然后在边缘似然的证据下界（ELBO）可以由下面的公式计算：

$$\log p_\theta(\mathcal{R})\geq \mathrm{ELBO}(\mathcal{R}) \tag{3-31}$$

$$=\mathbb{E}_{q_\varphi(\boldsymbol{Z}_1,\boldsymbol{Z}_2|\mathcal{R})}\log\left[\frac{p_\theta(\boldsymbol{Z}_1)p_\theta(\boldsymbol{Z}_2|\boldsymbol{Z}_1)p_\theta(\mathcal{R}|\boldsymbol{Z}_1,\boldsymbol{Z}_2)}{q_\varphi(\boldsymbol{Z}_2|\mathcal{R},\boldsymbol{Z}_1)q_\varphi(\boldsymbol{Z}_1|\mathcal{R})}\right] \tag{3-32}$$

$$=\mathbb{E}_{q_\varphi(\boldsymbol{Z}_1,\boldsymbol{Z}_2|\mathcal{R})}[\log p_\theta(\mathcal{R}|\boldsymbol{Z}_1,\boldsymbol{Z}_2)+\log p_\theta(\boldsymbol{Z}_2|\boldsymbol{Z}_1) \tag{3-33}$$

$$+\log p_\theta(\boldsymbol{Z}_1)-\log q_\varphi(\boldsymbol{Z}_2|\mathcal{R},\boldsymbol{Z}_1)-\log q_\varphi(\boldsymbol{Z}_1|\mathcal{R})] \tag{3-34}$$

$$=\mathbb{E}_{\boldsymbol{Z}_1\sim q_\varphi(\boldsymbol{Z}_1|\mathcal{R}),\boldsymbol{Z}_2\sim q_\varphi(\boldsymbol{Z}_2|\boldsymbol{Z}_1)}[\log p_\theta(\mathcal{R}|\boldsymbol{Z}_1,\boldsymbol{Z}_2)] \tag{3-35}$$

$$-\mathbb{D}_{\mathrm{KL}}(q_\varphi(\boldsymbol{Z}_2|\mathcal{R},\boldsymbol{Z}_1)||p_\theta(\boldsymbol{Z}_2|\boldsymbol{Z}_1))-\mathbb{D}_{\mathrm{KL}}(q_\varphi(\boldsymbol{Z}_1|\mathcal{R})||p_\theta(\boldsymbol{Z}_1)). \tag{3-36}$$

公式中的第一项表示重构成本（Reconstruction cost），它对观测扩散的负对数似然（Negative log-likelihood）进行估计，它激励模型去对来自于一系列隐藏变量 $\boldsymbol{Z}_1$ 和 $\boldsymbol{Z}_2$ 的序列的参与者进行有效率的解码。两个 DL 散度项 $\mathbb{D}_{\mathrm{KL}}(\cdot)$ 是模型的正则化项，它们鼓励模型去将推断的隐藏因子和两个先验进行匹配，两个先验分别为各向同性多元高斯（Isotropic multivariate Gaussian）和条件混合高斯（Conditional mixture of Gaussian）。这两个正则化项反应了在优化证据下界的时候所产生的信息损失（Information loss）。

3.5.3 使用正则化流的变分推断

在上一节中，我们在 CasFlow 模型中结合了低阶节点级别的变分自编码器和高阶级联级别的变分自编码器来对信息传播中的不确定性进行建模。特别地，对输入数据所学习到的隐藏表示均来自于两个变分自编码器所预先假设的简单的高斯后验分布族（Simple families of Gaussian posterior distribution）。但是，在实际的应用中，高斯分布的假设与现实生活中许多存在的更为复杂的分布相比过于简单，在表示数据分布时不够灵活。这一点在信息级联数据中尤为关键，因为它们的规

模分布是极度扭曲的[12,107]。因此，简单的高斯分布假设影响了变分推断的质量。为了能够使模型推断更为复杂、灵活、可扩展的后验分布族，我们使用了一个强大的基于概率的技术：正则化流（Normalizing Flows, NFs）[108–110]，来构建丰富的后验分布近似。

给定一个隐藏随机变量 $\boldsymbol{Z} \in \mathbb{R}^{dz}$（在本文中它是从高阶变分自编码器中学习到的 $\boldsymbol{Z}_2$），正则化流是一类生成模型，它将已观测到的向量 $\boldsymbol{Z}$ 变换为所需求的目标隐藏向量 $\boldsymbol{Z}^{(K)}$，该变换由一系列数量为 K 的可逆映射（Invertible mappings）组成，该变换的雅克比矩阵是可以计算且函数是可微的。更详细地说，正则化流使用了映射函数 $f: \boldsymbol{Z} \rightarrow \boldsymbol{Z}'$，其定义如下：

$$q(\boldsymbol{Z}') = q(\boldsymbol{Z}) \left| \det \frac{\partial f^{-1}}{\partial \boldsymbol{Z}'} \right| = q(\boldsymbol{Z}) \left| \det \frac{\partial f}{\partial \boldsymbol{Z}} \right|^{-1}, \tag{3-37}$$

其中 $q(\boldsymbol{Z})$ 是随机向量 $\boldsymbol{Z}$ 的分布，转换函数 f 是可逆的。为了从初始的密度 $q_0(\boldsymbol{Z})$ 中得到一个有效的概率密度 $q_K(\boldsymbol{Z}^{(K)})$，一系列 K 个正则化流的层级变换连续地使用公式3-37来计算目标密度：

$$\boldsymbol{Z}^{(K)} = f_K(\boldsymbol{Z}^{(K-1)}) = f_K(f_{K-1}(\ldots f_2(f_1(\boldsymbol{Z})))), \tag{3-38}$$

$$\ln q_K(\boldsymbol{Z}^{(K)}) = \ln q_0(\boldsymbol{Z}) - \sum_{k=1}^{K} \ln \det \left| \frac{\partial f_k}{\partial \boldsymbol{Z}^{(k)}} \right|. \tag{3-39}$$

如果该映射函数是恰当的，那么学习到的隐藏随机向量的混合分布与简单独立的高斯分布相比，前者更符合真实数据的分布。

为了更有效率地使用正则化流来对数据中的后验分布进行推断，考虑下面定义的变换：

$$f(\boldsymbol{Z}) = \boldsymbol{Z} + \boldsymbol{u}h(\boldsymbol{w}^T\boldsymbol{Z} + b), \tag{3-40}$$

其中 $\boldsymbol{w} \in \mathbb{R}^{dz}$，$\boldsymbol{u} \in \mathbb{R}^{dz}$ 和 $b \in \mathbb{R}$ 是它的参数，$h(\cdot)$ 是一个平滑的非线性函数。那么对数雅克比矩阵项（Logdet-Jacobian term，见公式3-37），近似的后验分布（Approximate posterior distribution，见公式3-39），以及边缘似然（Marginal

likelihood，见公式3-31）可以被改写为

$$\psi(\boldsymbol{Z}) = h'(\boldsymbol{w}^T\boldsymbol{Z} + b)\boldsymbol{w}, \tag{3-41}$$

$$\det\left|\frac{\partial f}{\partial \boldsymbol{Z}}\right| = \left|\det(\boldsymbol{I} + \boldsymbol{u}\psi(\boldsymbol{Z})^T)\right| = \left|1 + \boldsymbol{u}^T\psi(\boldsymbol{Z})\right|, \tag{3-42}$$

$$\ln q_K(\boldsymbol{Z}^{(\boldsymbol{K})}) = \ln q_0(\boldsymbol{Z}) - \sum_{k=1}^{K} \ln\left|1 + \boldsymbol{u}_k^T\psi_k(\boldsymbol{Z}^{(k)})\right|, \tag{3-43}$$

$$\log p_\theta(\mathcal{R}) \geq \mathrm{ELBO}(\mathcal{R}) + \mathbb{E}\left[\sum_{k=1}^{K} \ln\left|1 + \boldsymbol{u}_k^T\psi_k(\boldsymbol{Z}^{(k)})\right|\right]. \tag{3-44}$$

3.5.4 信息级联规模预测器

现在我们已经从双层双向门控循环单元中获得到了 $\boldsymbol{h}_2$，从层级变分自编码器和正则化流的一系列变换中获得到了 $\boldsymbol{Z}_3 = \boldsymbol{Z}^{(K)}$，我们使用下面的公式来将它们输入到多层感知机中来对最终的信息级联规模做出预测：

$$\hat{P}_k(t_p) = \mathrm{MLPs}(\mathrm{Concat}(\boldsymbol{h}_2, \boldsymbol{Z}_3)). \tag{3-45}$$

模型通过下面的均方误差损失函数来进行优化：

$$\mathcal{L}(\mathcal{R}_k;\Theta) = \frac{1}{N}\sum_{k=1}^{N}(\log\hat{P}_k(t_p) - \log P_k(t_p))^2 - \mathrm{ELBO}(\mathcal{R}_k), \tag{3-46}$$

其中 N 是所有信息级联的数量，$P_k(t_p)$ 是真实的信息级联规模（例如，转发级联 C_k 的用户数量），$\hat{P}_k(t_p)$ 是对信息级联 C_k 规模的预测，$\mathrm{ELBO}(\mathcal{R}_k)$ 是需要通过公式3-44来最大化的证据下界。

3.5.5 模型复杂度分析

本小节对模型的计算复杂度进行分析。因为信息级联的规模通常服从于厚尾分布族[12,89]，而且典型的在线社交网络通常拥有数百万个用户（也即是说全局图中的节点数和边数会非常大），如何有效率地对信息级联图和全局图进行计算是信息级联学习系统中一个非常关键的挑战。与传统的信息级联图学习模型相比，特别是那些基于随机游走的模型[65]和基于图神经网络的模型[76,77]，CasFlow 模型可以有效率地快速处理大规模的图数据，它对信息级联图和全局图的计算复杂度与图中边的数量成线性关系。

特别地，令 $|\mathcal{V}_c|$ 和 $|\mathcal{E}_c|$ 代表信息级联图中节点和边的数量，$|\mathcal{V}_g|$ 和 $|\mathcal{E}_g|$ 代表全局图中节点和边的数量，d_c 和 d_g 是信息级联图和全局图中节点的维度。

算法 3-1 CasFlow 模型的学习过程

输入： 观测到的信息级联 $C_k(t_o)$ 和它对应的全局图 $\mathcal{G}_g$。
输出： 预测的信息级联规模 $\hat{P}_k(t_p)$。

从 $C_k(t_o)$ 生成信息级联图 $\mathcal{G}_c$；
计算每一个节点 u_i 的图小波 $\Psi_{u,s}$；
计算信息级联图的节点嵌入 $\boldsymbol{E}_c(\mathcal{V}_c)$；
计算全局图的节点嵌入 $\boldsymbol{E}_g(\mathcal{V}_g)$；
while not 收敛 **do**
    训练双向门控循环单元来获取 $\boldsymbol{h}_2$；
    for 对 $|\mathcal{V}_c|$ 中的每一个用户 i **do**
        计算 $\boldsymbol{z}_i$；
    end for
    获取 $\vec{\boldsymbol{Z}}_1 = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_{|\mathcal{V}_c|}\}$；
    训练级联变分自编码器来获取 $\boldsymbol{Z}_2$；
    通过 K 次变换来获取 $\boldsymbol{Z}_3$；
    结合 $\boldsymbol{h}_2$ 和 $\boldsymbol{Z}_3$ 来做出最终预测；
end while

3.5.5.1 计算信息级联图中节点嵌入的复杂度

图谱小波是由切比雪夫多项式（Chebyshev polynomials）[111] 来计算的，它的时间复杂度是 $O(h|\mathcal{E}_c|)$，与图中的边的数量成线性关系，其中 h 是切比雪夫多项式近似的阶数 [112]。

3.5.5.2 计算全局图中节点嵌入的复杂度

如 [98] 所述，计算截断奇异值分解和 QR 分解的复杂度是 $O(d_g^2|\mathcal{V}_g|)$，因为 $d_g \ll |\mathcal{V}_g|$，所以稀疏矩阵分解的总体复杂度是 $O(d_g^2|\mathcal{V}_g| + |\mathcal{E}_g|)$。

3.5.5.3 CasFlow 模型中其他部分的复杂度

门控循环单元和多层感知机的时间复杂度和空间复杂度与隐藏变量的维度有关。在我们的实验设置中，CasFlow 拥有大约两百万个可训练的参数，在 batch 大小为 64 的情况下，CasFlow 模型需要花费大约 83 微秒来进行单步训练，花费大约 6.78 分钟来生成全局图 $\mathcal{G}_g$ 的节点嵌入，该全局图中大约有一千五百万条边。我们将在下一节中对 CasFlow 模型和基准模型的效率进行对比。CasFlow 模型的总体学习过程请参见算法3-1。

3.6 实验

在本节中，我们首先对信息级联数据集进行介绍，然后介绍最先进的基准模型、实验结果、消融实验、以及对模型的各种分析。

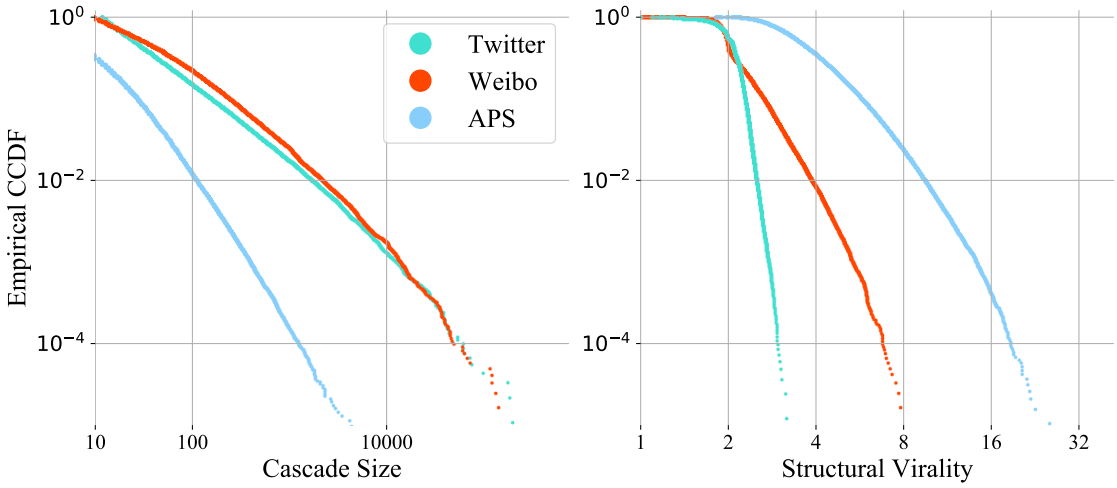


图 3-2 三个数据集上信息级联规模和结构扩散性（Structural virality）的经验互补累积分布函数（Empirical Complementary Cumulative Distribution Function, CCDF）。推特和微博的信息级联规模一般要大于 APS 论文的规模，但是它们的级联扩散性很低。

3.6.1 实验设置

3.6.1.1 数据集

信息级联的类型有多种，例如，社交网络中的推特和微博，图片、音乐、视频等多媒体，邮件，新闻文章，研究论文的引用等，都可以形成信息级联。我们选择了三个常见的公开数据集：推特、微博、APS，它们在已有的信息级联规模预测模型中广泛使用[21,55,76]。

推特数据集（Twitter）由 [87] 的作者们收集，包括发布于 2012 年 3 月 24 日到 4 月 25 日之间在推特平台上发布的由英文书写的推特。我们将标签和它们的参与者作为独立的信息级联。推特数据集的全局图由多种关系组成，包括用户之间相互的关注和被关注关系、转发关系、提醒关系等。信息级联图由上述所有的三种关系构成。

新浪微博（Sina Weibo）是中国最大的微博平台。微博数据集中的微博及其转发微博可以构成转发信息级联[21]。微博数据集中的全局图由所有用户之间的转发关系构成。

APS 数据集包含发布于美国物理学会所承办的各类期刊上的科学论文。APS 数据集中每一篇论文及其引用论文构成了引用信息级联。APS 数据集中的全局图由所有论文的作者之间的引用关系所构成。

从三种截然不同的数据集上进行实验可以考察模型在没有特征工程和领域知

表 3-2 三个信息级联数据集的统计信息

数据集	推特	微博	APS
信息级联数量	88,440	119,313	207,685
全局图中的节点数	490,474	6,738,040	616,316
全局图中的边数	1,903,230	15,249,636	3,304,400
平均规模	142	240	51
观察时间内信息级联的数量			
训练集（1 天/0.5 小时/3 年）	9,639	21,463	18,511
验证集（1 天/0.5 小时/3 年）	2,066	4,599	3,967
测试集（1 天/0.5 小时/3 年）	2,065	4,599	3,966
训练集（2 天/1 小时/5 年）	12,739	29,908	32,102
验证集（2 天/1 小时/5 年）	2,730	6,409	6,879
训练集（2 天/1 小时/5 年）	2,729	6,408	6,879
信息级联图统计信息			
平均序列长度	2.196	2.237	3.999
平均结构扩散性	1.995	2.025	3.114
平均 Page Rank 值	0.073	0.045	0.189
平均图密度	0.183	0.090	0.310

识加持之下的泛化性能。数据集的详细统计信息请见表3-2和图3-2。我们可以观察到，相比推特和微博数据集，APS 拥有较小的平均规模，但是后者的结构扩散性（通过计算维纳系数 Wiener index 得到[12,113]）要更高，这预示着科研论文的规模主要是通过其他论文来进行传播扩散的。对于推特标签来说，它们的结构扩散性一般来说最小，这说明它的信息传播主要依靠自身广播。

另外，依据已有工作的设定[21,76]，对于推特数据集，观察时间被设置为 1 天和 2 天，预测时间为 32 天；对于微博数据集，观察时间被设置为 0.5 小时和 1 小时，预测时间为 24 小时；对于 APS 数据集，观察时间被设置为 3 年和 5 年，预测时间为 20 年。我们剔除了在观察时间内信息级联规模小于 10 的信息级联。对于观察时间内信息级联规模大于 100 的信息级联，我们只选取其前 100 个参与者。对于推特标签，我们选取发表在 4 月 10 日之前的标签，从而保证数据集中的标签均有超过 15 天的增长时间。对于微博数据集，从用户每天的作息习惯考虑，我们选取早上八点到晚上六点之间的微博，这个设定这使得每个选中的微博均有超过六个小时以上的时间去增长规模。对于 APS 数据集，我们考虑发表在 1893 年到 1997 年之间的论文，这样使得每篇论文至少拥有 20 年（1997 年到 2017 年）的时间来增长引用。

3.6.1.2 基准模型

为了验证我们提出的 CasFlow 模型在信息级联规模预测上的有效性，我们选取了三类基准模型来与 CasFlow 模型进行对比。

- **基于特征工程的模型**：是一类最常见的信息级联预测模型。这些模型首先从数据中人工抽取特征，然后将这些特征输入到机器学习模型之中来进行训练和验证。例如，Szabo 等人[26] 使用观测到的信息级联规模 $P_j(t_o)$ 来预测新闻文章和在线视频的规模 $\hat{P}_j(t_p)$。这个模型使用了观测到的信息级联规模和累积规模序列作为特征，我们称其为 Feature-S&H。Cheng 等人[12] 使用了五种类型的特征来预测信息级联的增长，包括内容特征、原始内容发布者及转发者的特征、结构和时序特征。特别地，这些特征包括累积规模序列、原始发布时间和第一个参与者到达时间的差、所有参与者时间的前一半的平均、所有参与者时间的后一半的平均、叶子节点的数量、平均节点度、平均和最大序列长度等。我们将这些特征输入到线性回归模型和多层神经网络中，这两个模型分别被称为 Feature-Linear 和 Feature-Deep。
- **统计生成模型**：是一类基于时间序列的信息级联预测模型，例如 Pinto 等人[88] 提出的时间序列模型（我们称其为 TimeSeries）。Cao 等人[21] 提出的模型结合了深度学习和霍克斯点过程来预测信息级联规模。它包括了三个霍克斯点过程的关键元素，也就是用户的影响力、自激励机制、以及时间衰减因素。我们将这个模型称为 DeepHawkes。
- **基于深度学习的模型**：CasCN[76] 是一个基于图卷积网络（GCN）的信息级联预测框架，它同时对时间和结构信息进行建模。它对信息级联图进行了子图采样，然后使用了长短期记忆网络（LSTM）来捕获信息级联的增长过程。DMT-LIC[93] 是一个多任务学习模型，它通过一个共享参数层和注意力/门控机制来同时对用户级别的行为和级联级别的预测进行建模。需要注意的是，我们没有同一些预测模型（例如 DeepCas[65]、CYAN-RNN[22]、DeepInf[19]、FOREST [92] 等）进行相比，因为它们主要对微观的节点激活行为进行预测而不是信息级联规模预测，或者它们只考虑到了信息级联中的结构信息。

3.6.1.3 参数设置

对于三个数据集，我们随机将其分割为训练集（70%）、验证集（15%）、测试集（15%）。对于所有的模型包括 CasFlow，通过早停策略在训练集上训练到最好的效果（验证集损失连续 10 轮不下降）。对于基准模型，学习率和 L_2 系数从 $10^{\{0,-1,-2,\ldots,-8\}}$ 中进行选择。DeepHawkes、CasCN、DMT-LIC 的节点嵌入维度设置

表 3-3 CasFlow 模型与基准模型在推特数据集上的对比。观察时间为 1 天和 2 天，评价指标为均方对数误差（MSLE）和平均绝对百分比误差（MAPE），指标数值越低代表效果越好。我们还对预测结果做了配对 t-检验，$*$ 符号表示与基准模型相比，CasFlow 的提升具有统计上的显著性，$p < 0.001$。

模型	推特			
	1 天		2 天	
	MSLE	**MAPE**	**MSLE**	**MAPE**
Feature-S&H	14.792	0.960	13.515	0.983
TimeSeries	8.214	0.547	6.023	0.445
Feature-Linear	9.326	0.520	6.758	0.459
Feature-Deep	7.438	0.485	6.357	0.500
DeepHawkes	7.216	0.587	5.788	0.536
CasCN	7.183	0.547	5.561	0.525
DMT-LIC	7.152	0.467	5.427	0.481
CasFlow-LocalStruct	7.254	0.475	5.366	0.370
CasFlow-GlobalStruct	11.244	0.704	10.619	0.709
CasFlow-Temporal	7.258	0.450	5.436	0.375
CasFlow-Structural	10.860	0.680	9.927	0.620
CasFlow-RNN	7.273	0.467	5.392	0.377
CasFlow-VAE	7.138	**0.428**	5.178	**0.337**
CasFlow*	7.340	0.435	5.119	0.383
CasFlow-noNF	7.272	0.429	**5.083**	0.345
CasFlow	**6.954***	0.455*	5.143*	0.361*
(效果提升)	↑2.7%	↑8.3%	↑6.3%	↑24.3%

为 50，批大小设置为 64，所有的其他超参数被设置为原始值。

对于 CasFlow 模型中对信息级联图进行嵌入时使用到的尺度参数 s，我们使用一种经过理论验证的方法 [96] 来选择 s 所在的合适范围 $[s_{\min}, s_{\max}]$。这也就是说，我们直接使用 $s_{\min}$ 和 $s_{\max}$ 来生成最终的节点嵌入 $E_c(u_i) = [E_{c,s_{\min}}(u_i), E_{c,s_{\max}}(u_i)]$。当 $d = 10$ 的时候（均匀分布的点），嵌入的维度 d_c 为 40。对于全局图中的节点嵌入 $\boldsymbol{E}_g(\mathcal{V}_g)$，嵌入维度 d_g 设置为 40。隐藏因子 $\boldsymbol{Z}_1$、$\boldsymbol{Z}_2$ 和 $\boldsymbol{Z}_3$ 的维度均设置为 64。门控循环单元的隐藏单元数量为 128，正则化流的变换次数 K 为 8。多层感知机的隐藏单元数量分别为 64 和 32。

表 3-4 CasFlow 模型与基准模型在微博数据集上的对比。观察时间为 1 天和 2 天，评价指标为均方对数误差（MSLE）和平均绝对百分比误差（MAPE），指标数值越低代表效果越好。我们还对预测结果做了配对 t-检验，$*$ 符号表示与基准模型相比，CasFlow 的提升具有统计上的显著性，$p < 0.001$。

模型	微博			
	0.5 小时		1 小时	
	MSLE	MAPE	MSLE	MAPE
Feature-S&H	4.455	0.390	4.001	0.398
TimeSeries	3.119	0.277	2.693	0.268
Feature-Linear	2.959	0.258	2.640	0.271
Feature-Deep	2.715	0.228	2.546	0.272
DeepHawkes	2.891	0.268	2.796	0.282
CasCN	2.804	0.254	2.732	0.273
DMT-LIC	2.752	0.249	2.689	0.270
CasFlow-LocalStruct	2.681	0.228	2.488	0.251
CasFlow-GlobalStruct	3.014	0.274	2.780	0.291
CasFlow-Temporal	2.691	0.228	2.566	0.272
CasFlow-Structural	2.939	0.266	2.797	0.292
CasFlow-RNN	2.444	0.217	2.234	**0.232**
CasFlow-VAE	2.712	0.260	2.561	0.272
CasFlow*	2.429	0.217	**2.206**	0.245
CasFlow-noNF	2.501	0.223	2.291	0.246
CasFlow	**2.402***	**0.210***	2.279*	0.238*
(效果提升)	↑12.7%	↑15.7%	↑18.0%	↑13.4%

3.6.1.4 评价指标

参考之前的工作[10,21]，我们使用均方对数误差（Mean Square Logarithmic Error, MSLE）和平均绝对百分比误差（Mean Absolute Percentage Error, MAPE）来对预测的结果进行评价，它们具体的定义如下：

$$\mathrm{MSLE} = \frac{1}{N}\sum_{i=1}^{N}(\log_2 \Delta\hat{P}_i - \log_2 \Delta P_i)^2, \tag{3-47}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\log_2 \Delta\hat{P}_i - \log_2 \Delta P_i|}{\log_2 \Delta P_i}, \tag{3-48}$$

其中 N 是测试集中信息级联的数量，$\Delta P_i = P_i(t_p) - P_i(t_o)$ 是信息级联规模的增量。

除了这两个评价指标，我们还采用了测定系数（Coefficient of determination,

表 3-5 CasFlow 模型与基准模型在 APS 数据集上的对比。观察时间为 1 天和 2 天，评价指标为均方对数误差（MSLE）和平均绝对百分比误差（MAPE），指标数值越低代表效果越好。我们还对预测结果做了配对 t-检验，$*$ 符号表示与基准模型相比，CasFlow 的提升具有统计上的显著性，$p < 0.001$。

模型	APS			
	3 年		5 年	
	MSLE	**MAPE**	**MSLE**	**MAPE**
Feature-S&H	2.382	0.316	2.348	0.350
TimeSeries	1.867	0.271	1.735	0.291
Feature-Linear	1.852	0.272	1.728	0.291
Feature-Deep	1.844	0.270	1.666	0.282
DeepHawkes	1.573	0.271	**1.314**	0.335
CasCN	1.562	0.268	1.421	0.265
DMT-LIC	1.539	0.264	1.398	0.258
CasFlow-LocalStruct	1.814	0.267	1.686	0.285
CasFlow-GlobalStruct	1.478	0.241	1.546	0.266
CasFlow-Temporal	1.798	0.266	1.682	0.283
CasFlow-Structural	1.480	0.237	1.574	0.273
CasFlow-RNN	1.367	0.227	1.365	**0.244**
CasFlow-VAE	1.463	0.234	1.481	0.271
CasFlow*	**1.346**	0.223	1.373	0.251
CasFlow-noNF	1.370	0.227	1.401	0.251
CasFlow	1.361*	**0.222***	1.354*	0.248*
(效果提升)	↑12.5%	↑15.9%	↓-2.2%	↑5.4%

R^2）和 Top-k 覆盖百分比（COV-k）来作为评价指标，后者由对于最大的 k 个信息级联中预测正确的比率所定义，也就是说，对于 k 个最大的信息级联，以及模型预测后认为最大的 k 个级联，它们两个集合的交集除以 k。在本节中，我们将 k 的值设置为 $k = \lfloor N/10 \rfloor$。

3.6.1.5 实验环境

本节的实验环境为：Intel E5-2680 v4 2.40GHZ 处理器，单个 NVIDIA GeForce GTX 1080Ti 显卡，64GB 内存。CasFlow 模型由 TensorFlow 框架搭建，使用了 Adam 优化器，CasFlow 所有的时间消耗（包括预处理、训练、测试等）小于一个小时。

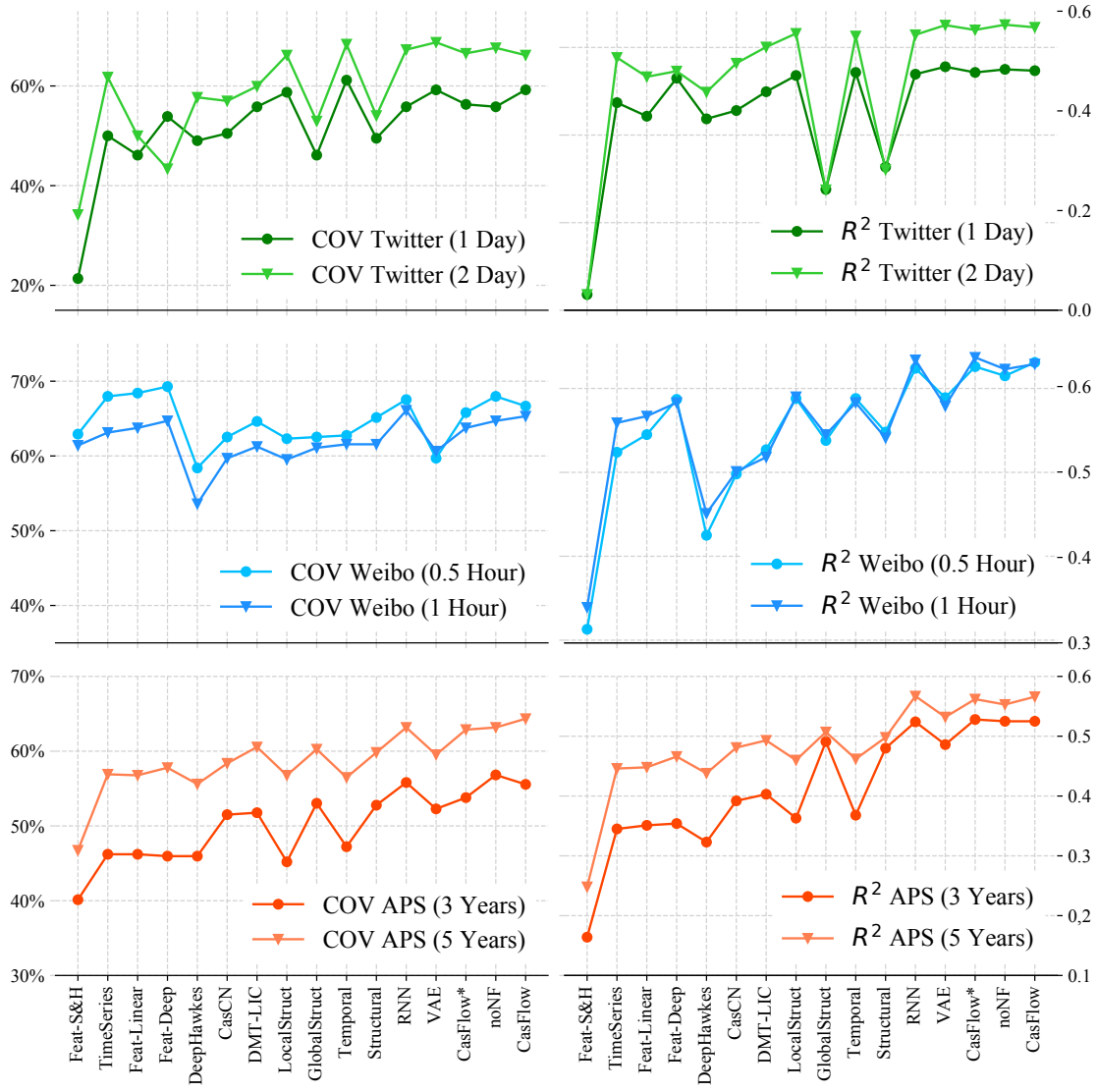


图 3-3 CasFlow 模型与基准模型在三个数据集上的效果对比。评价指标是 Top-10% 覆盖（COV）和测定系数（R^2）。

3.6.2 效果对比

CasFlow 模型的预测效果，以及和基准模型的对比结果，请见表3-3（推特）、表3-4（微博）和表3-5（APS），以及图3-3。我们拥有以下发现：

发现一　CasFlow 模型显著地提升了基准模型的预测效果。在 Weibo 数据集上，CasFlow 模型超过了基准模型中表现最好的 DMT-LIC 多达 15.2%，这一结果表明 CasFlow 的层级信息级联建模可以有效地提升预测效果。

发现二　基于特征工程的模型和其他基准模型的差距并不大，在某些情况下，

特征工程模型和生成模型的效果甚至超过了基于深度学习的模型，这一结果表明深度学习模型并不总是比特征工程模型好。但是，特征工程模型的效果取决于专家设计的特征，设计这些特征需要特定的专业领域知识，而且设计的特征泛化性能也不高。例如，在 APS 数据集上，基于特征工程的模型预测效果非常差。

发现三　另一方面，因为 DeepHawkes 模型没有考虑到信息级联中的拓扑结构，所以它的效果依赖于时间序列建模能力和传播路径。此外，由于 DeepHawkes 模型使用了自激励机制，使得它倾向于高估信息级联的规模。CasCN 考虑了信息级联中的时间和结构特征，但是它只考虑了局部结构学习，忽略了全局用户行为。

发现四　在所有的基准模型中 DMT-LIC 表现最好，它的多任务学习机制不仅考虑到了信息级联中的结构特性，还考虑到了节点的个体行为特征。在某种程度上，它隐含地学习了信息级联中的层级信息。CasFlow 相比 DMT-LIC 的效果提升在于前者对信息传播中的不确定性进行了建模，并且同时考虑了节点级别和级联级别的不确定性。

为了更好地探究 CasFlow 模型中每一个模块对预测结果的贡献，我们设计和实现了以下八个 CasFlow 模型的变种：

（1）CasFlow-LocalStruct：我们移除了 CasFlow 模型中全局节点嵌入 $\boldsymbol{E}_g(\mathcal{V}_g)$ 模块。

（2）CasFlow-GlobalStruct：我们移除了 CasFlow 模型中局部节点嵌入 $\boldsymbol{E}_c(\mathcal{V}_c)$ 模块。

（3）CasFlow-Temporal：我们移除了 CasFlow 模型中的结构建模模块。对于 CasFlow-Temporal 来说，信息级联图中的所有的非根节点直接连接到了根节点，并且我们不使用全局图信息。

（4）CasFlow-Structural：我们移除了 CasFlow 模型中的时序建模模块。

（5）CasFlow-RNN：我们只使用双层双向门控循环单元输出的 $\boldsymbol{h}_2$ 来对信息级联进行建模和预测，也即是说，这个变种里没有层级变分推断模块。

（6）CasFlow-VAE：我们只使用了层级变分自编码器和正则化流输出的 $\boldsymbol{Z}$ 来对信息级联进行建模和预测，也即是说，这个变种里没有双层双向门控循环单元。

（7）CasFlow*：是一个浅层版本的 CasFlow 模型，它只使用了低阶的不确定性建模和循环神经网络来对信息级联做出预测，即 $\boldsymbol{Z}_1$ 和 $\boldsymbol{h}_2$。

（8）CasFlow-noNF：我们移除了正则化流部分，也就是说，我们使用 $\boldsymbol{Z}_2$ 和 $\boldsymbol{h}_2$ 来

对信息级联做出预测。

在表3-3、表3-4和表3-5，以及在图3-3中，我们可以看到 CasFlow 模型和它的变种模型的效果对比。对比结果显示：（1）在推特和微博数据集上，CasFlow-LocalStruct 的预测效果显著地高于 CasFlow-GlobalStruct，但是在 APS 数据集上，它们的预测效果有着明显的下降，这说明在 APS 数据集上全局结构信息可能是更有用的特征；（2）CasFlow-Temporal 在推特和微博数据集上的效果均比 CasFlow-Structural 要好。但是，结果显示，对于 APS 论文级联来说，结构信息要更重要。这一现象表明，同时对时序信息和结构信息建模可以使模型的泛化能力得到提升；（3）出乎意料的是，在没有对层级级联效果进行建模的情况下，CasFlow-RNN 取得了非常好的效果，这可能归功于其强大的节点嵌入模块，同时考虑了局部和全局结构信息。此外，CasFlow-VAE 在推特数据集上表现的非常好，这表明了其对传播不确定性进行建模的优点；（4）CasFlow* 和 CasFlo-noNF 相比其他模型，取得了有竞争力或者更好的效果，这表明对层级传播不确定性和更复杂的后验分布进行建模的重要性。

3.6.3 模型可解释性

在本小节中，我们从以下四个角度来模型预测的可解释性进行分析：（1）使用 *t*-SNE 对学习到的信息级联表示进行可视化；（2）对 CasFlow 模型中的重要超参数进行敏感度分析；（3）分析 CasFlow 模型的运行时间，并与其他基准模型进行对比；以及（4）对信息级联规模的分布进行分析。

3.6.3.1 隐藏表示

为了对 CasFlow 模型所学习到的信息级联表示有一个直观的理解（特别是对于变分自编码器和正则化流的部分），依据之前的工作[65,76]，我们使用 *t*-SNE 技术[114] 来绘制 CasFlow 模型学习到的信息级联高阶表示的 2D 可视化图像（如图3-4所示）。这些信息级联表示来自于 CasFlow-RNN、CasFlow-noNF，以及完整的 CasFlow 模型。

这里我们使用的数据集是微博，观测时间为 0.5 小时。图中的每一个点代表测试集中的一个信息级联。测试集中共有 4,599 个信息级联。点的颜色越深，意味着它所对应的值越高。图中共有五种类型的值，从上到下依次为信息级联规模（a-c）、结构扩散性（d-f）、第一次转发时间 t_1（g-i）、边密度（j-l）以及平均反应时间（m-o）。图中每一行代表一种类型。图中的每一列表示不同隐藏表示，第一列的表示（a, d, g, j, m）来自于 CasFlow-RNN 中多层感知机中的最后一层，第二列的表示（b, e, h, k, n）来自于 CasFlow-noNF 的 $\boldsymbol{Z}_2$，第三列的表示（c, f, i, l, o）

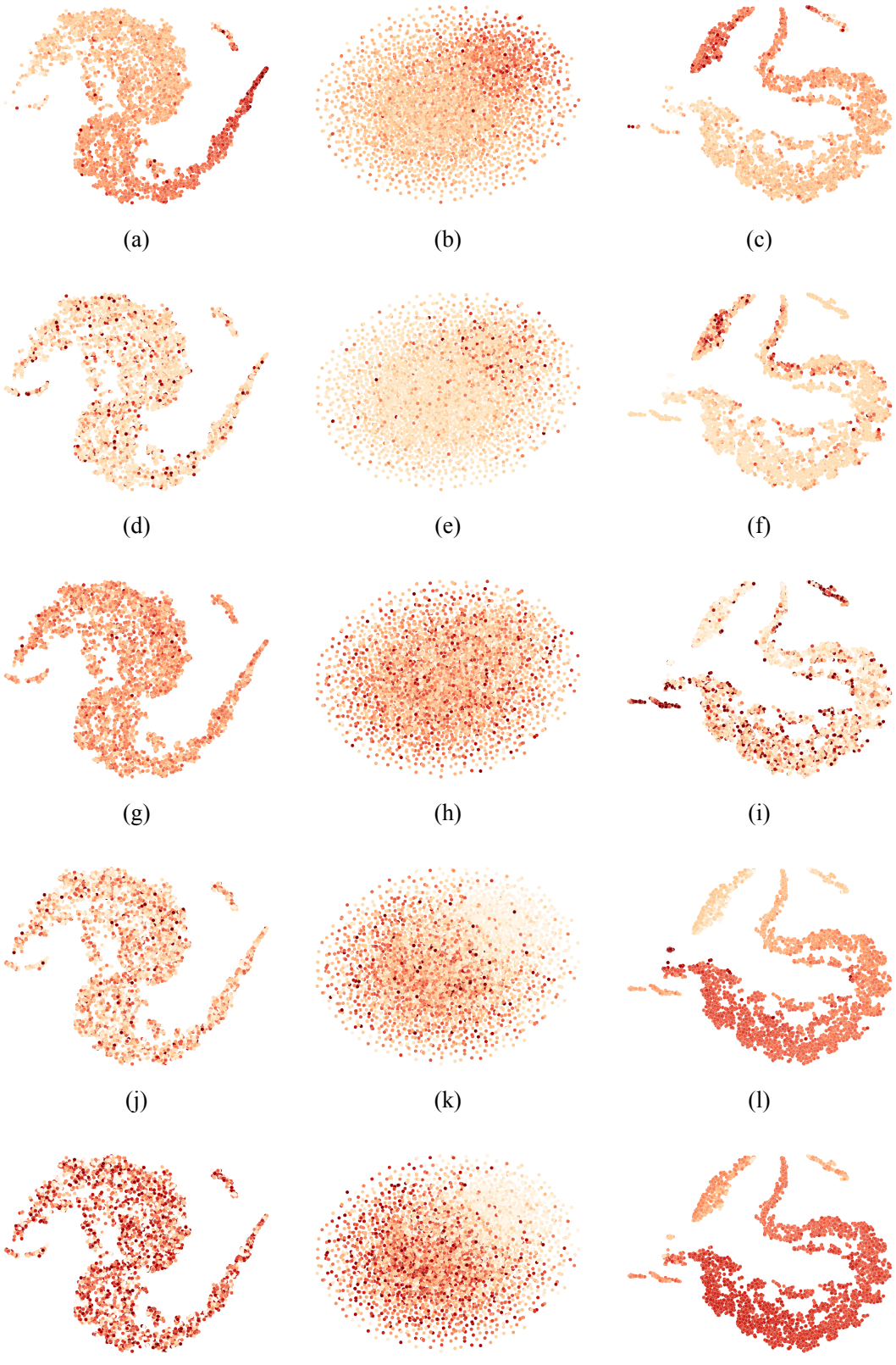


图 3-4 在微博数据集上学习到的隐藏表示的 *t*-SNE 可视化。

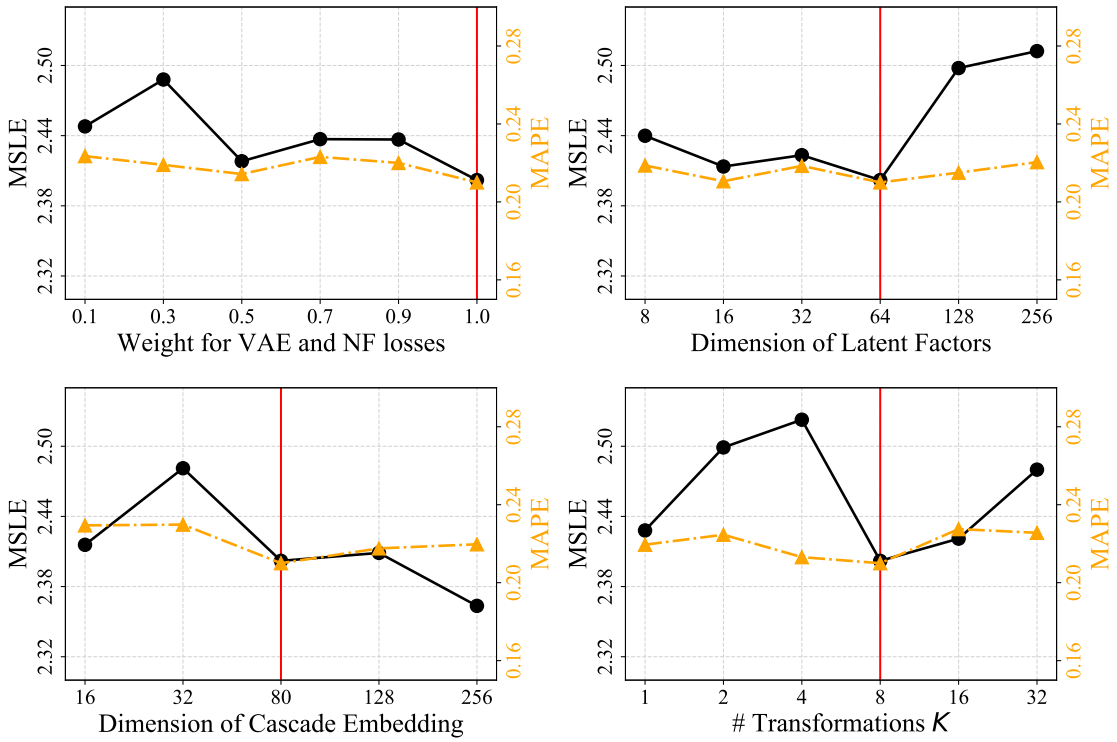


图 3-5 CasFlow 模型中四种重要的超参数对模型预测效果的影响分析。使用的数据集是微博，观测时间是 0.5 小时，评价指标为均方对数误差（MSLE）和平均绝对百分比误差（MAPE）。垂直线的值代表本章中实验的默认设置。

来自于 CasFlow 中多层感知机中的最后一层。

如图3-4的第一列所示，除了信息级联的规模特征，其他的特征，不管是基于结构特征还是时间特征，并没有显示出明显的与信息级联规模有关联的特性。这一现象表明，如果只使用循环神经网络来对信息级联进行建模，模型无法对信息级联的特征和学习到的隐藏表示关联起来。相反，在模型的第二列中，信息级联的表示是 CasFlow-noNF 模型中层级变分自编码器的输出 $\boldsymbol{Z}_2$，图像中信息级联点的分布符合于变分自编码器的高斯假设。从图中我们可以明显地看出，除了第一次转发时间，其他的特征显示出了明显的聚类效果，这些聚类与信息级联的规模密切相关，例如，对于大的结构扩散度，小的边密度和平均转发时间，信息级联的规模倾向于更大。值得注意是，在 CasFlow 中我们并没有使用这些特征来进行训练或者测试，但是模型自身学习到了与未来规模有关联的、有意义的、可解释的表示。最终，如图的最后一列所示，相比 CasFlow-RNN 和 CasFlow-noNF，完整的 CasFlow 模型不仅在预测效果上达到了最佳，它的可解释性也更好。例如图3-4(i)中，规模大的信息级联通常与小的第一次转发时间 t_1 相关联。以上实验说明了在信息级联规模预测中集成变分推断和正则化流的优点。

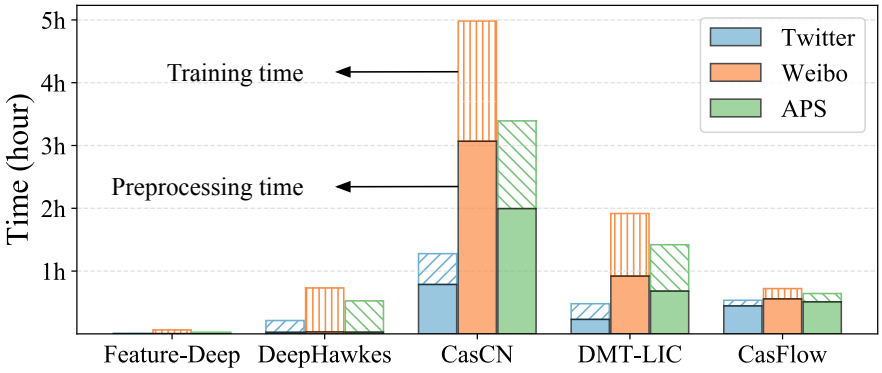


图 3-6 CasFlow 模型在预处理阶段和训练阶段的计算时间消耗。这里对比了四个基准模型（Feature-Deep、DeepHawkes、CasCN、DMT-LIC），使用了所有三个数据集，预测时间分别为：1 天（推特）、0.5 小时（微博）、3 年（APS）。

3.6.3.2 超参数对模型的影响

CasFlow 模型中有多个超参数（Hyper-parameters），它们数值的选择可能对模型的预测效果造成很大的影响。我们在微博数据集上对 CasFlow 模型中的四个重要超参数（变分自编码器和正则化流损失函数的权重、隐藏因子的维度、信息级联嵌入的维度、正则化流变换的次数 K）进行了超参数敏感度实验。图3-5中垂直线的值代表本章中实验的默认参数值，图中左坐标轴显示了预测的均方对数误差（MSLE）值，右坐标轴显示了预测的平均绝对百分比误差（MAPE）值。详细的实验结果解释如下：

- 变分自编码器和正则化流损失函数权重的影响：我们对变分自编码器和正则化流损失函数的权重进行了调整，权重的大小表示了信息级联监督训练学习和信息传播不确定性学习之间的一种权衡。
- 隐藏因子和信息级联嵌入的维度：我们对隐藏因子维度 d_Z 和信息级联嵌入维度 $d_c + d_g$ 尝试了不同的值（范围从 8 到 256）。从图中可知，大的嵌入维度有时候会让预测的效果下降。
- 正则化流变换的次数：我们尝试了不同的正则化流变换次数（从 1 到 32），我们观察到 8 次变换取得了最好的预测效果。

3.6.3.3 时间消耗

我们计算了 CasFlow 模型预处理阶段和训练阶段所需要的消耗的时间，包括四个基准模型（Feature-Deep、DeepHawkes、CasCN、DMT-LIC）。从图3-6可知，CasFlow 模型与 CasCN 和 DMT-LIC 模型相比更为效率，与 DeepHawkes 的效率处于同一水平。

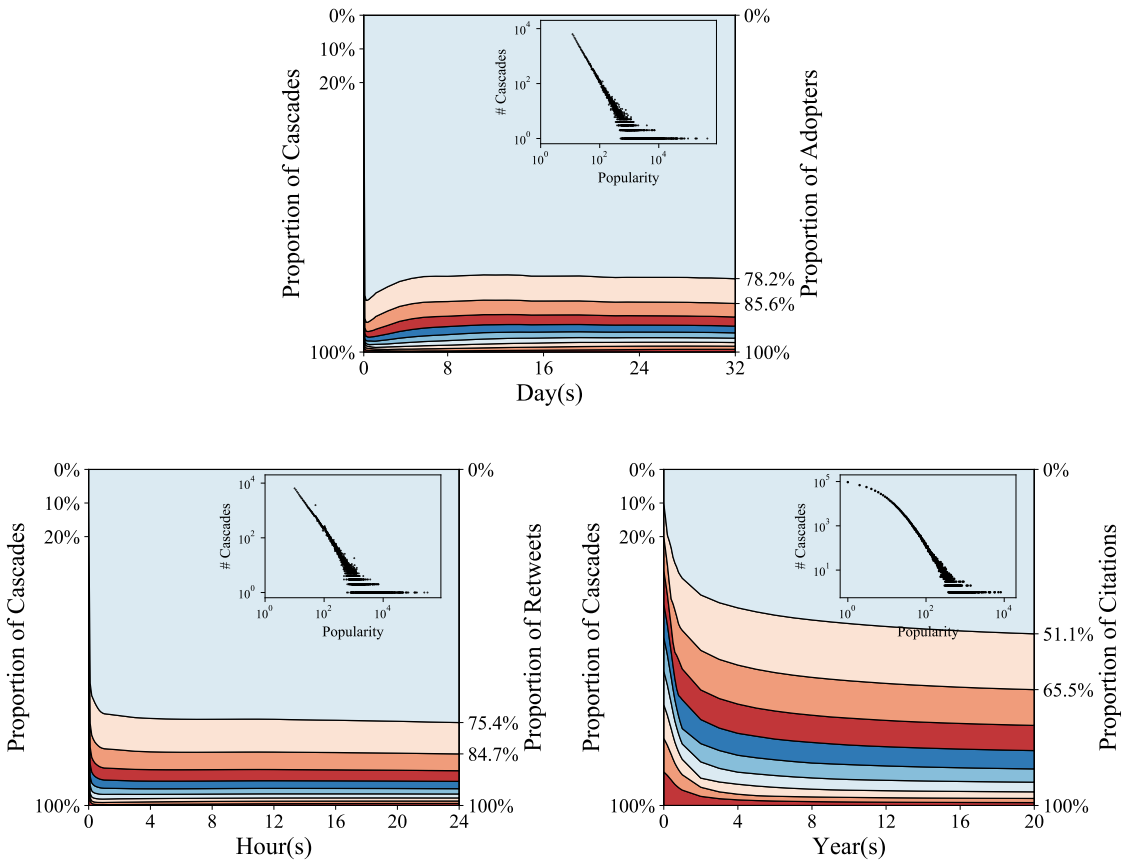


图 3-7 对于三个数据集中信息级联的参与者数量随着时间变化的堆叠绘制图。右上角的嵌入图：信息级联规模的分布，它们都符合幂律分布（Power-law distribution）。

3.6.3.4 信息级联规模分布

最后，我们来探究不同类型的信息级联的规模分布，这可以让我们对信息传播的内在机制有更多的理解。图3-7中显示了三个信息级联数据集的规模分布。

对于推特和微博数据集来说（两个都是社交媒体网站），最流行的 20% 的标签/微博，在发布之后不久，就拥有超过 80% 的参与者，这符合帕累托分布（Pareto distribution），也就是经典的二八定律。对于 APS 数据集来说，情况是相似的，其展现出一种更缓和的信息垄断，例如，对于所有论文发表的前二十年中，大约 50% 的引用来自于最有影响力的 10% 的论文。此外，三个数据集中的衰减趋势并不相同，例如，最流行的标签和微博在发布之后的短暂时间内就占据了大部分的关注，对于引用最多的一系列论文，它们的影响力是随着时间逐渐增长的。因此，如何确定有影响力的信息，以及如何对不同类型的信息判别其时间衰减因素，是一个重要的研究问题。

3.7 本章小结

在本章中，我们提出了 CasFlow 模型，它是第一个基于贝叶斯学习和图表示学习的深度信息级联预测模型。CasFlow 模型利用了层级变分自编码器来对信息传播中的不确定性进行建模，并且使用了正则化流来学习更为丰富和灵活的信息级联后验分布。我们的实验证明，在三个大规模的信息级联数据集上，与当前最先进的基准预测模型相比，CasFlow 模型显著地提升了它们的预测效果，并且提供了一定程度上的预测可解释性。总的来说，我们的发现证明，使用深度生成模型来对信息级联数据进行训练和优化是一个有希望的探索方向。

第四章 基于图对比自监督学习的信息级联规模预测

本章主要介绍图对比自监督学习的信息级联规模预测模型 CCGL 的实现细节以及相关实验。本章首先讨论与 CCGL 模型有关的相关工作，并解释了 CCGL 模型如何解决已有模型所面临的挑战。CCGL 模型从对比的、自监督的、以及不基于特定任务的角度来学习信息级联图的表示。它首先设计了一个效率的信息级联图数据增强策略 AugSIM 来捕捉信息传播中的变化和不确定性。然后同时利用增强后的有标签数据和无标签数据来进行对比自监督预训练，计算正样本和负样本之间的距离（相似度），然后对正样本对进行匹配，从而学习到通用的、鲁棒性高的信息级联图表示。CCGL 在下游任务上进行基于特定任务的模型微调和知识蒸馏来提升预测效果。随后我们介绍了 CCGL 与互信息最大化之间的关联，以及 CCGL 的计算复杂度。最后我们介绍了文本对 CCGL 模型所做的实验内容，包括数据集处理、基准模型（包括监督模型、半监督模型，以及其他的数据增强策略）、实验参数设置、实验结果等。我们还对 CCGL 模型做了大量的消融实验、参数敏感度实验以及迁移学习实验。

实验结果表明，CCGL 模型取得了信息级联图预测的当前最佳效果。我们还有以下有意义的发现：（1）CCGL 很大程度上克服了预测上的过拟合问题，在所有五个只带有少量标签的信息级联数据集上，CCGL 显著地提升了预测效果，例如，在只有 1% 的标签数据的情况下，微调和知识蒸馏过后的 CCGL 模型最多降低了 9.5% 的信息级联规模预测误差，并且最多提升了 19.9% 的信息级联爆发预测准确率；（2）对于信息级联对比自监督学习模型来说，大的模型和深的映射网络是非常重要的，而增加的 batch 大小和训练轮数并不能带来额外的效果提升；（3）一个使用了无标签数据的教师学生知识蒸馏网络可以显著地减轻知识迁移的副作用；（4）在不同信息级联数据集和不同的信息级联预测任务上的知识迁移学习表明，CCGL 显著地超过了有监督学习模型的预测效果。图4-1对 CCGL 模型和传统的监督模型进行了对比。表4-1中介绍了本章中常用的数学符号。

4.1 相关工作介绍

随着深度学习的流行，越来越多的信息级联预测模型采用了深度学习技术来对信息级联中的各种特性进行建模。这些模型[21]需要大量的标注数据（有些时候可能会非常难以获取）来进行有监督训练（Supervised training）。这些模型在不同的数据集之间，或者在不同的现实预测任务上也难以直接应用，泛化性能较

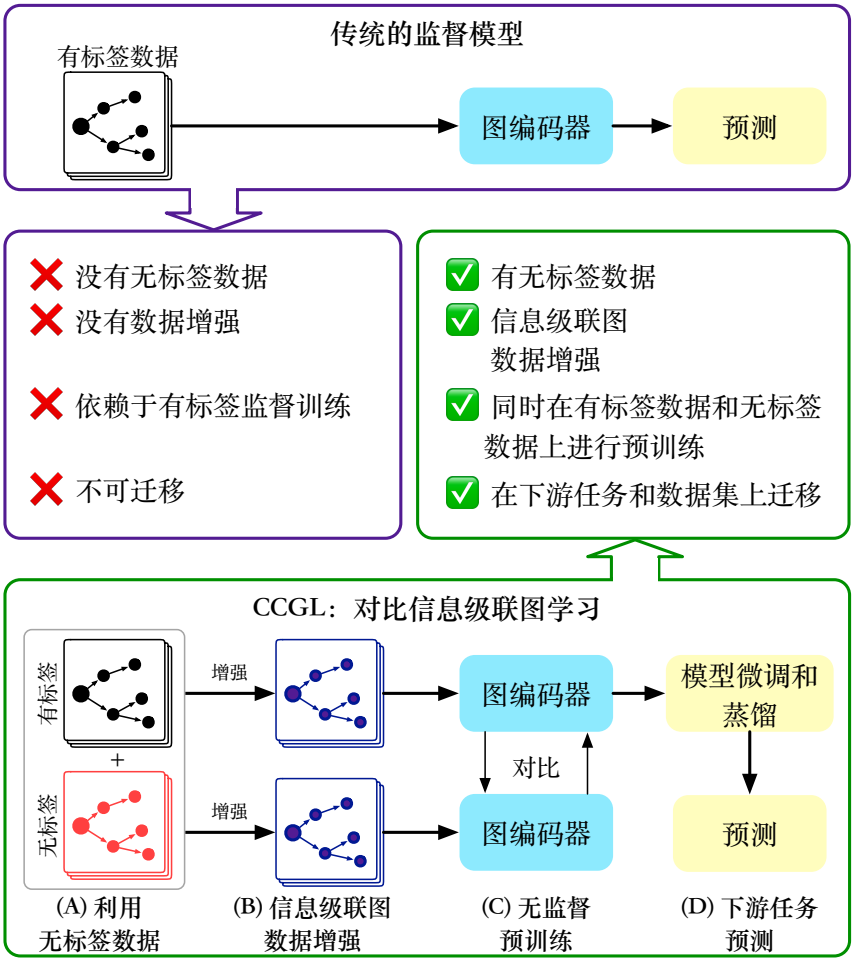


图 4-1 CCGL 模型与传统的监督模型的对比

弱。此外，在现实中，还有很多无标签数据的存在，有监督训练模型很难直接对这些数据进行利用。为了利用大量的无标签数据来学习到更通用的数据表示，研究者们转而开始探究各种半监督（Semi-supervised）或者无监督（Unsupervised）模型，例如各类生成式模型（Generative models），包括无监督领域适应图卷积网络（Unsuper vised domain adaptive graph convolutional network）[115]、自编码变分贝叶斯（Auto-encoding variational bayes）[102]、以及对图神经网络进行预训练的各种策略[116]。这些模型使用细粒度的低阶嵌入特征去重构样本，容易导致在特定数据上的过拟合[21]。在许多情况下（例如信息级联预测），我们需要学习到更为鲁棒和通用的表示，这些表示可以很好地在嵌入空间中对不同的样本进行分辨。也就是说，学习一种更为高阶的、抽象的语义表示会对信息级联预测更为有效。这一点启发了我们使用去采用对比自监督学习（Contrastive self-supervised learning）。这一种学习范式在近些年来获得了成功，特别是在自然语言处理（Natural Language Processing）和计算机视觉（Computer Vision）领域。对比自监督学习模型特别适

表 4-1 本章中所用到的数学符号及其描述

符号	描述
a_j^i	信息级联 C_i 中节点 u_j^i 的吸引力（Attractiveness）。
B	批大小（Batch size）。
C_i	信息级联。
d	表示的维度。
$\mathcal{E}_i, \mathcal{V}_i$	信息级联图 $\mathcal{G}_i$ 中边和节点的集合。
$\mathcal{G}_i, \tilde{\mathcal{G}}_i$	信息级联图和数据增强之后的信息级联图。
$\boldsymbol{h}_i$	在下游任务上使用的隐藏表示。
$\mathcal{L}$	对比损失函数。
M	级联中用户（节点）的数量。
N	有标签信息级联图的数量。
$\mathcal{N}(u_j^i)$	节点 u_j^i 的邻居数量。
$P_i(t_p)$	信息级联 C_i 在时间 t_p 的预测。
r_j^i	信息级联 C_i 中节点 u_j^i 的移除概率。
s	随机游走重启概率。
t_j^i	用户 u_j^i 的参与时间。
u_j^i	信息级联 C_i 中的用户。
U	无标签信息级联的数量。
$\boldsymbol{z}_i$	计算对比损失函数时所用到的隐藏表示。
η_i	数据增强力度。
λ	指数分布的速率参数。
τ	对比损失函数的温度参数。

合在无人类标注的无标签数据集上进行学习，它提升了模型的泛化性能和迁移性能。但是，图数据上的对比自监督训练还是一个相对较新的领域，特别是还没有人研究过使用对比自监督学习的图上的信息传播建模。

4.1.1 信息级联建模

信息级联建模在信息传播领和社交网络分析领域中是一个广泛研究的和关键的问题[11]。大多数已有的工作可以分为三种类型：（1）基于特征工程的模型学习信息级联中所蕴含的各种人工设计的特征，例如时序特征、结构特征、文本和图像特征、信息级联元数据特征和历史行为等；（2）生成式的时序模型主要学习时间序列数据，它们使用了各种随机过程来对网络中的信息传播进行基于生成式的和统计的建模[10,117]；（3）基于深度学习的模型主要使用了各类先进的神经网络技术来学习信息级联的有效表示[77,78]。基于深度学习的模型中有一些使用了图表示学习和图神经网络，例如，DeepCas[65] 对信息级联图进行了建模，它使用了多个

随机游走过程；CasCN [76] 使用了图卷积网络（GCN）来捕获信息级联图中的结构信息。尽管这些模型取得了一定的成功，它们依赖于有监督训练，泛化性能较弱。

4.1.2 自监督学习

自监督学习（Self-supervised learning）使用了数据本身作为训练信号，从而可以从大量的无标签数据中进行有效的表示学习，提升学习到的表示的质量，进而提升下游任务上的预测效果。已有的自监督学习模型通常使用了一种对比学习（Contrastive learning）的范式，对比学习通过对比正负样本来学习数据的表示。它们主要关注于前置任务（Pretext tasks）和对比损失函数（Contrastive losses）的设计：

- 前置任务的类型包括：从噪音数据中恢复原始数据、预测邻近的单词、或者是对原始数据的转换等。它们的例子包括正负样本之间的实例判别（Instance discrimination）[118,119]、全局-局部对比的相对位置预测 [120]、或者互信息最大化模型（Mutual Information Maximization）[121–123] 等。这些前置任务被用来学习和抽取有用的数据表示，但是前置任务并不是真实的需要的预测目标。
- 对比损失函数关注于正负样本之间的相似度，它对于学习到好的表示至关重要。负样本采样策略是对比损失函数设计中最重要的模块之一。之前的工作设计了许多学习策略，例如，端到端学习（End-to-end learning）[121]、记忆机制（Memory mechanism）[118]、映射网络（Projection head）、数据增强和模型微调（Data augmentation and model fine-tuning）等 [124,125]。

4.1.3 图上的数据增强

对比自监督学习中一个非常重要的模块是数据增强，用来对模型的泛化性能和预测效果进行提升。数据增强被广泛地应用在自然语言处理和计算机视觉模型中 [126]。之前的自监督学习模型使用了各种数据增强策略 [120,124]。与文本和图像不同，图数据通常是非欧式的（non-Eucliean）、稀疏的、复杂的。图像可以是有向图、动态图、属性图、甚至是异构图等，这一系列特性增加了对图数据进行建模的复杂程度。因为图数据的不规则性，语言和图像数据上的增强策略一般不能直接利用到图数据中，图数据上专用的数据增强策略也较少被研究。如何设计有效的图数据增强策略是一个有挑战的重要任务，可以用来从图结构中使用无监督学习的范式来抽取通用的知识。已有的图像数据增强方法，例如旋转、随机裁剪和调整尺寸、颜色扰动、高斯模糊等策略 [125]，不能直接利用到图数据上来进行预训练和模型微调 [116]。

对图数据进行增强最直接的方法是对其进行增加或者删除节点和边。不过，

这样的操作可能会面临着各种困难[127]，例如：如何选择增加的或者删除的节点和边；如何对新增加的节点进行标注；如何处理与节点或边有关联的属性；等等。已有的图数据增强工作包括：随机移除边和覆盖边的部分属性[116,128]来避免过拟合（over-fitting）和过平滑（over-smoothing）；使用模型的预测[129]来选择图神经网络中增加或者删除的边并以此来对抗过平滑问题；使用图自编码器（Graph auto-encoder）[127]来作为边预测模块，来对图增加“缺失”的边，以及删除“噪音”边。

最近，Qiu 等人[130]提出了 GCC 模型，它使用了子图实例判别（Sub-graph instance discrimination）来作为预训练任务。一个 r 自我网络的数据增强过程分为三步：带有重新启动的随机游走（Random walk with restart, RWR），子图归纳（Sub-graph induction），以及匿名化（Anonymization），来对图数据结构中通用和可迁移的模式进行捕捉。Hassani 等人[131]提出了一种增强图数据的策略，它分为两个部分：（1）操纵节点的特征，例如，遮盖节点属性，或者对其添加高斯噪音；（2）操纵图的结构，例如，增加或删除连结，或者使用子采样（Sub-sampling）。

在本章中，我们设计了创新的信息级联图数据增强策略。首先，我们使用了无标签数据来进行模型预训练，使其学习到信息级联的有效表示。然后，我们定义了信息级联图数据增强器，它模仿了信息在图中的传播方式。我们提出的数据增强策略有三个独有的特性：（1）它适用于信息级联图，而大多数之前的策略不能直接使用到信息级联图中；（2）我们对图中的节点和边，包括节点和边的属性，同时进行了操纵；（3）我们模型的目的是信息级联预测，而不是节点分类或图分类。

4.1.4 图上的预训练模型和迁移学习

无监督的图预训练模型的一个重要的应用场景是预训练到一个可迁移的图编码器，然后将其应用到下游基于图的任务上。迁移学习在视觉学习和语言学习领域中已经较为常见了，但是解决图上的预训练和迁移任务的模型还很少[116,132]，如何在图数据上设计图预训练策略和缓解负面转移（Negative transfer）是非常有挑战性的课题。Hu 等人[116]提出了一个基于图神经网络的策略，该策略整合了节点级别和图级别的预训练，从而可以在迁移的时候处理分布之外（Out-of-distribution）的样本。UDA-GCN 模型[115]模型在图知识迁移学习的时候同时考虑了局部和全局一致性。GPT-GNN 模型[133]是另外一个生成式的基于图神经网络的预训练模型，它使用了属性和边生成器来进行在大规模图上的自监督预训练。

在本章中，CCGL 模型极大地扩充了训练数据（包括有标签数据、无标签数据），从不同的数据领域对信息级联图进行了增强，从而使模型在大量的数据里学

习到通用和鲁棒的图表示，然后将学习到的知识迁移到下游预测任务上（对数据稀少的下游任务非常有用）。CCGL 不依赖于特定的前置任务或者领域知识，它可以被扩展到其他类型的信息级联图预测任务上。

4.2 CCGL 模型总体框架

在本节中，我们介绍 CCGL 模型的总体框架。模型的框架见图4-2，实现代码见https://github.com/Xovee/ccgl。它主要包含以下三个部分：

（1） **信息级联图数据增强**：为了从信息级联图表示中学习到更为通用的和可迁移的知识，我们设计了一种利用了无标签数据和有标签数据的数据增强策略，它通过模拟信息在网络中扩散的行为来捕捉信息传播中的变化和不确定性。

（2） **对比自监督预训练**：我们使用了对比预训练框架来学习信息级联图的抽象级别的表示，从而解决许多其他生成式的表示学习框架容易面临的过拟合现象。

（3） **模型微调和知识蒸馏**：对于特定的下游任务，我们将预训练好的模型进行在有标签数据上的微调。我们还设计了知识蒸馏网络来使训练好的模型更为鲁棒，泛化性能更好。如果没有知识蒸馏网络对模型的优化训练，基于特定任务的模型在其他任务或数据集上进行迁移时，容易遇到负面迁移问题，进而导致预测效果的下降。

为了更好地探索和理解一个无监督学习框架是如何提升信息级联图表示学习的能力，我们首先关注于回答以下三个科学问题：

问题一 **无监督数据会不会提升信息级联表示学习的能力和信息级联预测？** 大多数的监督学习模型无法利用无标签数据。在基于图的信息级联学习中，无标签的信息级联图可以从它们的早期进化阶段进行构造。在传统的信息级联预测模型中，这些信息级联图被简单地排除在训练之外[10,21,78]。自监督学习模型关注于从大规模的无标签数据中用一种无监督的和不基于特定任务的方式来学习到数据的有效表示。无标签数据可以从一个数据集[124]或者多个数据集中获取[130]。我们相信，在模型中结合无标签数据可以提升学习到的信息级联表示的鲁棒性和泛化性。

问题二 **图数据增强会不会提升信息级联预测的效果？如果是的话，我们该如何设计有效的信息级联图数据增强策略？** 因为信息级联图数据结构的特殊性，文本或图像的数据增强策略不能直接在信息级联图上

使用。如何设计对信息级联图的新的数据增强策略是提升信息级联预测效果和提升模型泛化能力的一个非常重要的挑战，这需要我们从图学习和对比学习两个角度进行考虑。如我们之前所讨论的，之前的图数据增强技术主要关注于图神经网络、节点或图分类任务[128,129,134]，它们中的大部分模型也只关注于边和属性的操纵，忽略了对节点的处理。这启发了我们去设计对信息级联图进行图数据增强的策略。在 CCGL 中，我们提出了一个创新的信息级联图增强策略 AugSIM。我们通过模拟信息在网络中传播的机制来对图中的每一个节点按照它们的到达时间顺序进行遍历，然后对每一个节点计算一个基于节点度的吸引力概率，根据这个概率，图中的节点可以吸引新的参与者，或者失去已有的参与者。

问题三 **对比自监督学习框架会不会提升信息级联预测效果？**对比学习最近在语言和视觉任务上取得了非常大的成功。但是，据我们所知，还没有工作对信息级联上的对比学习进行过研究。在本章中，基于迁移学习的观点[122,125,134]，我们设计了 CCGL 模型，它支持各种图编码网络，图表示学习模型和图神经网络，或者专注于信息级联学习的模型（例如 DeepCas[65]、VaCas[78]、Coupled-GNNs[77] 等），它们都可以当做 CCGL 模型中的图编码器。我们在 CCGL 模型中对各种自监督学习技术进行了实现和对比，目标在于探寻无监督信息级联模型的学习能力，以及寻求提升下游信息级联预测任务效果的办法。

4.3 使用无标签信息级联图

在典型的信息级联预测任务上，无标签数据一般被简单地排除在训练和测试之外。以预测信息级联的规模为例，对于一条微博或者论文，我们首先观察其初期的增长数据，然后预测其未来某个时间的规模。在 [21] 中，对于一个科研论文，作者们首先观察其前几年的引用情况，然后预测其在第 20 年的引用数量。但是，对于发表时间不足 20 年的论文，也就是说，在这种设定下，这些论文没有合适的标签，所以就被排除在数据集之外。如图4-3所示，对于 APS 数据集中的论文，只有大约 43.3% 带有标签的论文被利用了，剩下的无标签论文没有参与到训练和测试的过程中。这也就说，这一个预测系统只能利用发表时间超过 20 年的论文数据。

如果我们使用这个预测系统来预测新发表论文的引用数量，我们会偏向于用旧的论文引用增长特点来预测新发表的论文，忽略了最近发表论文的传播特点和传播机制。为了解决这个问题，我们需要使用无监督学习的技术来将无标签的信

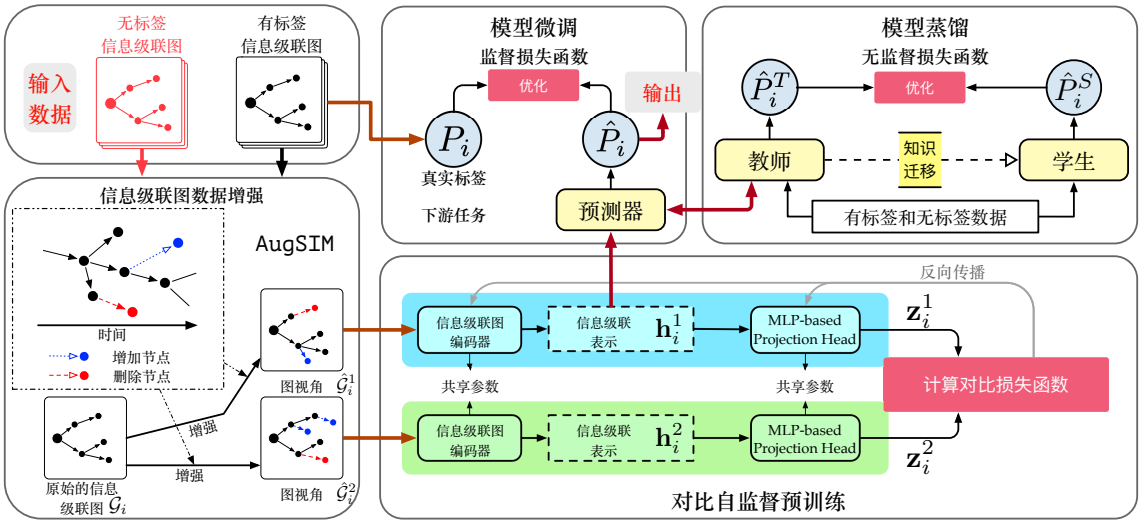


图 4-2 我们在本章中提出的信息级联图表示学习 CCGL 模型。它由三个主要的部分组成：（1）为信息级联图设计的数据增强策略 AugSIM；（2）基于非特定任务的无监督预训练模型，使用对比损失函数来优化模型；（3）模型微调和知识蒸馏网络。预训练阶段和蒸馏阶段同时利用了有标签数据和无标签数据。

息级联数据纳入考虑。

4.4 信息级联图数据增强

已有的自监督学习框架[118, 124, 130]所用到的数据增强策略是为语言模型、视觉模型、以及基于图神经网络的模型所设计的。它们不能直接应用到信息级联图之中，主要是由于以下几个原因：（1）当前并没有直观的方法将文本或图像数据增强策略直接应用到信息级联图上；（2）信息级联图是树结构的数据，从一个根节点开始（例如一个新发布的微博或论文），不断地增长、扩散到网络的其他部分，收获更多的关注和影响力（例如转发和引用）。对信息级联图进行任意的增加节点、删除节点、增加边、删除边，会极大地破坏信息级联图的主体结构；（3）信息级联图中的节点附有时序特征，即节点的到达时间，而时序特征被认为是信息级联预测中至关重要的特征之一。为了解决这三个挑战，我们提出了一个创新的信息级联图数据增强策略：AugSIM.

4.4.1 AugSIM：模拟信息传播过程的信息级联图数据增强

为了给目标信息级联图创建不同的图视角，我们设计了一种基于用户影响力和参与时间的数据增强策略，该策略背后的思想简单，运行效率快，并且捕获到了一定程度的图拓扑结构或节点属性之间的相似程度。创建的图视角可以在随后

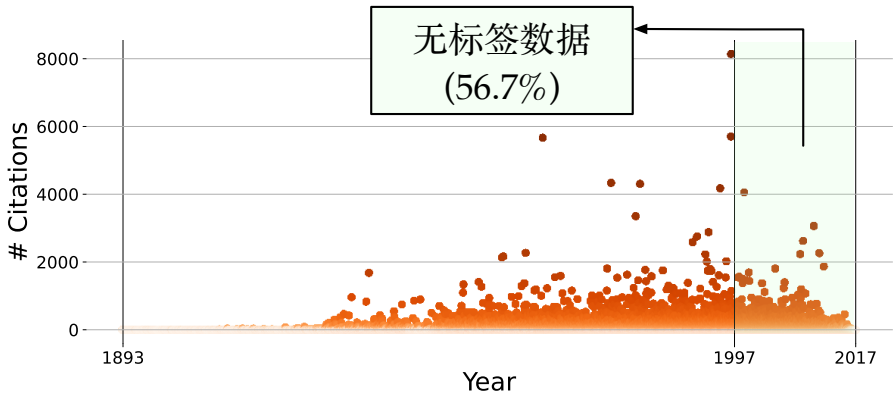


图 4-3 APS 数据集中从 1993 年 6 月到 2017 年 12 月的所有 616,316 篇论文。在预测的目标设置为 20 年之后的引用时，发表日期晚于 1997 年 12 月的论文没有合适的标签。

的对比学习框架中使用。

对于信息级联图 $\mathcal{G}_i$ 中的每一个用户 u_j^i，我们对其计算一个吸引力概率（Attractiveness probability）a_j^i 来管理节点增加过程：

$$a_j^i = \eta_i \frac{\text{degree}(u_j^i)}{\sum_{u_k^i \in \mathcal{V}} \text{degree}(u_k^i)}, \tag{4-1}$$

其中增强力度 η_i 是一个级联级别的超参数，它控制着往信息级联图 $\mathcal{G}_i$ 中增加的节点的数量。增加的节点 u_{new}^i 与 u_j^i 相连，并且我们给它赋予一个节点属性：参与者时间 $t_{\text{new}}^i \in [t_j, t_o]$。参与者时间 t_j 可以被当做是人类反应时间的一个实例[62]。我们同时计算局部（级联级别）的参与者时间 t_{local}^i 和一个全局（数据集级别）的参与者时间 t_{global} 来对新增加的节点赋予新的参与者时间：

$$t_{\text{new}}^i = t_j + \theta_t t_{\text{local}}^i + (1-\theta_t) t_{\text{global}}, \tag{4-2}$$

其中 θ_t 是一个权值参数，它控制着两种参与者时间的权重，t_{local}^i 是信息级联 C_i 中所以参与者时间的平均 $(\frac{1}{|\mathcal{V}_i|}\sum_{j\in|\mathcal{V}_i|} t_j)$。$t_{\text{global}}$ 是一个全局参与者时间，它来自于一个指数分布：

$$f(t;\lambda) \sim \lambda e^{-\lambda t}, \quad \text{with} \quad \lambda > 0. \tag{4-3}$$

在上述公式4-3中，λ 是一个速率参数，它可以从数据集中所有的参与者时间中进行经验拟合。在对信息级联图中的每一个节点进行遍历之后，我们对该图添加了一系列节点 $\mathcal{V}_i^{\text{new}} = \{u_{\text{new},k}^i | k = 1, 2, \dots\}$、相对应的节点特征 $\{t_{\text{new},k}^i | k = 1, 2, \dots\}$，以及相对应的边 $\mathcal{E}_i^{\text{new}}$，添加的节点直接连接到它的双亲节点上。在 $\mathcal{V}_i^{\text{new}}$ 中添加的

节点的期望数量由 η_i 控制，即 $E(|\mathcal{V}_i^{\text{new}}|) = \sum_{j=1}^{|\mathcal{V}_i|} a_j^i = \eta_i$。

我们接下来讨论如何去移除信息级联图中的节点和边。类似地，我们对扩充后的节点集合 $\mathcal{V}_i \cup \mathcal{V}_i^{\text{new}}$ 进行遍历。对于其中的每一个叶子结点 $u_j^i \in \mathcal{V}_i^{\text{leaf}}$，我们计算一个移除概率（Removal probability）r_j^i，它定义为：

$$r_j^i = \eta_i \frac{\text{degree}(\text{parent}(u_j^i))}{\sum_{u_k^i \in \mathcal{V}_i^{\text{leaf}}} \text{degree}(\text{parent}(u_k^i))}, \tag{4-4}$$

其中 v_j^i 是 $\mathcal{V}_i^{\text{leaf}}$ 集合中 u_j^i 的双亲节点。移除的节点数量的期望为 $\sum_{j=1}^{\mathcal{V}_i^{\text{leaf}}} r_j^i = \eta_i$。算法4-1中详细描述了AugSIM数据增强策略。

为了简单起见，节点移除过程的对象只选取了叶子节点，这样的话，信息级联图的主体结构不会受到影响。其他更为复杂的方法也可以被用来模拟信息在网络中传播的过程，例如：（1）允许添加的节点再去吸引新的节点；（2）不仅移除叶子节点，也可以移除它们的双亲节点；（3）在增加或者删除节点、边、节点属性的时候，考虑到更多的信息级联特征，例如关注者和被关注者的数量、引用的数量、作者的 h 指数等，作为用户影响力的一种表现；（4）随机点过程，例如泊松过程和霍克斯自激励过程等（它们经常被用来对信息传播中的级联行为进行建模），将这些过程当做生成模型来对信息级联图进行增强（增删节点或边），或者用来对数据集进行扩充[133]。这些方法均可能提升信息级联图的数据增强过程，我们将其作为 CCGL 的未来工作。在本章中，我们使用节点的度和参与时间来增强信息级联图，创建相似但不同的信息级联图视角。这些信息级联图视角可以用在随后的对比自监督预训练之中。AugSIM策略可以被看做是信息在网络中的一次再传播（Re-diffusion），再传播保留了信息传播的基本特征，但是带来了一定程度上的变化和不确定性。

4.5 基于对比自监督学习的信息级联图表示学习

在上一节中，我们定义好了信息级联图数据增强，我们现在介绍 CCGL 模型是如何从无标签数据中进行通用表示学习的。

4.5.1 数据增强

我们首先使用数据增强策略AugSIM来对同一个信息级联图创建不同的视角。给定信息级联图 $\mathcal{G}_i$，我们首先对其进行两次数据增强操作，来创建两个不同但相似的视角，我们把它们称作 $\tilde{\mathcal{G}}_i^1$ 和 $\tilde{\mathcal{G}}_i^2$。这两个增强后的信息级联图被当做是一对正样本 $(\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2)$ 来参与后续的对比自监督预训练。

算法 4-1 CCGL 模型中的信息级联图数据增强策略

输入： N 个信息级联 $\{C_i\}_{i\in[1,N]}$ 和相对应的信息级联图 $\{\mathcal{G}_i=(\mathcal{V}_i,\mathcal{E}_i)\}_{i\in[1,N]}$.
输出： 增强过后的信息级联图 $\{\tilde{\mathcal{G}}_i\}_{i\in[1,N]}$.

定义数据增强过程：AugSIM
通过公式4-3计算全局参与者时间 t_{global}
for all $i\in\{1,\ldots,N\}$ **do**
  计算局部参与者时间 $t^i_{\text{local}}=\frac{1}{|\mathcal{V}_i|}\sum_{j\in|\mathcal{V}_i|}t_j$
  for all $j\in[1,|\mathcal{V}_i|]$ **do**
    通过公式4-1计算吸引力概率 a^i_j
    $\mathcal{V}^{\text{new}}_i=\emptyset,\mathcal{E}^{\text{new}}_i=\emptyset$
    if $\text{random}(0,1)\le a^i_j$ **then**
      从指数分布中采样 t_{global}
      $t^i_{\text{new}}=t_j+\theta t^i_{\text{local}}+(1-\theta)t_{\text{global}}$
      # 增加节点属性
      $\mathcal{V}^{\text{new}}_i=\mathcal{V}^{\text{new}}_i\cup\{u^i_{\text{new}}\}$
      # 增加节点
      $\mathcal{E}^{\text{new}}_i=\mathcal{E}^{\text{new}}_i\cup\{(u^i_j,u^i_{\text{new}})\}$
      # 增加边
    end if
  end for
  $\mathcal{V}_i=\mathcal{V}_i\cup\mathcal{V}^{\text{new}}_i,\mathcal{E}_i=\mathcal{E}_i\cup\mathcal{E}^{\text{new}}_i$
  for all $j\in[1,|\mathcal{V}^{\text{leaf}}_i|]$ **do**
    通过公式4-4计算移除概率 r^i_j
    if $\text{random}(0,1)\le r^i_j$ **then**
      $\mathcal{V}_i=\mathcal{V}_i\setminus\{u_j\}$
      # 移除节点及节点属性
      $\mathcal{E}_i=\mathcal{E}_i\setminus\{(u^{\text{parent}}_j,u_j)\}$
      # 移除边
    end if
  end for
end for

4.5.2 信息级联图建模

然后我们使用使用一个编码器将信息级联图编码为一个向量，该向量中蕴含了信息级联图中的时间和结构特征。图编码器的选择在这里是通用不受限制的，任何可以将稀疏的信息级联图编码为表示向量的网络都可以使用。我们采用了一个先进的信息级联预测模型 VaCas[78]，它主要包含两个部分：（1）一个基于图谱小波的图嵌入模块；（2）一个基于双向门控循环单元的模块。需要注意的是，该模型等价于 [78] 中描述的 Cas-RNN 模型。这两个模块将信息级联图 $\mathcal{G}_i$ 编码为一个固定长度的表示 $\boldsymbol{h}_i\in\mathbb{R}^{d_h}$。为了更好地理解隐藏因子，以及避免表示中可能出现的噪音数据，我们在网络之后添加了一个基于多层感知机的映射网络（Projection

head）[124]，来将表示 $\boldsymbol{h}_i$ 映射为一个新的表示 $\boldsymbol{z}_i \in \mathbb{R}^{d_z}$。之前的工作中[125,135]表明，这一映射网络可以提升学习到的表示的质量，进而提升下游任务上的预测效果。我们在本章的实验部分测试了不同的映射网络结构选择。映射网络定义如下：

$$\boldsymbol{h}_i^1 = \text{cascade\_encoder}(\tilde{\mathcal{G}}_i^1), \qquad \boldsymbol{z}_i^1 = \text{MLP}(\boldsymbol{h}_i^1), \tag{4-5}$$

$$\boldsymbol{h}_i^2 = \text{cascade\_encoder}(\tilde{\mathcal{G}}_i^2), \qquad \boldsymbol{z}_i^2 = \text{MLP}(\boldsymbol{h}_i^2). \tag{4-6}$$

需要注意的是，映射网络只参与到无监督学习阶段，也就是说，只有表示 $\boldsymbol{h}_i$ 参与到了下游任务的微调阶段。对于表示 $\boldsymbol{z}_i$，它只被用来计算对比损失函数，从而对 CCGL 模型进行优化。

4.5.3 对比损失函数

为了训练 CCGL 模型，我们使用对比损失函数[124]来优化模型，它定义为最大化相同信息级联图的两个增强视角之间的相似性，然后在一个 batch 中对正样本对 $(\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2)$ 和其他所有的负样本进行判别。特别地，CCGL 模型首先随机采样 B 个信息级联图，然后对每个信息级联图增强两次来获得 $2B$ 个增强后的信息级联图。对于一个 batch 中的正样本对，我们将剩下的 $2B-2$ 个信息级联图看作是负样本。给定一个相似度函数 $\text{sim}(\cdot,\cdot)$ 来度量两个向量之间的相似程度（我们使用了余弦相似度），则相对于正样本对 $(\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2)$ 来说，对比损失函数正式定义为：

$$\mathcal{L}_{1,2}^{\text{contrastive}} = -\log \frac{\exp\left(\text{sim}(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2)/\tau\right)}{\sum_{k=1}^{2B} 1(k \neq i) \exp\left(\text{sim}(\boldsymbol{z}_i^1, \boldsymbol{z}_i^k)/\tau\right))}, \tag{4-7}$$

其中 $1(\cdot)$ 是一个指示函数，τ 是一个温度函数。这个对比损失函数通常被称为 InfoNCE[123] 或 XT-Xent[124]，它被广泛地在自监督学习模型之中[118,119,130]。

4.5.4 损失函数讨论

通常对比学习需要选取很多的负样本来与正样本进行对比。但是受限于计算资源和经济预算，维护一个非常大的 batch/负样本字典，或者构建一个非常大的网络架构有些时候并不现实。例如，SimCLR 模型[124]在一个拥有 128 个 TPU v3 核心的平台上使用了大小为 8,192 的 batch（其中有 16,382 个负样本）。端到端模型的学习能力受限于 batch 的大小。有些模型依赖于特殊的前置任务，这可能导致网络结构的变化，例如限制感受野的大小[121]、将图分割为子图结构[134]等，限制了它们的泛化能力。记忆银行（Memory bank）[119]和动量更新（Momentum update）[118]是另外一类对比机制，它们对 batch 的大小和负样本的数量进行了解耦，并且提供了平滑和一致的编码器更新。之前的工作讨论了端到端模型和记忆机制的影

响，例如，在 [118] 中，作者们发现端到端模型在 batch 不大的情况下也具有有竞争力的性能。在 [130] 和 [125] 中，作者们总结到，在 batch 很大的情况下，记忆机制只提供了微小的提升。大的 batch 也需要更多的 GPU/TPU 内存和更长的训练时间。

4.6 在下游信息级联预测任务上进行模型微调和知识蒸馏

在经过通用的对比学习预训练之后，我们需要使用有标签数据在具体的下游任务上进行模型微调。在本文中，我们主要关注于信息级联规模预测问题。

4.6.1 模型微调

CCGL 模型的无监督预训练阶段同时利用了有标签数据和无标签数据。在基于特定下游任务的训练阶段，我们使用有标签数据来对模型进行微调（Fine-tuning）。在这个阶段，对比学习时用到的映射网路可以被完全丢弃（即只有信息级联图编码器会参与微调），部分丢弃（即编码器和部分映射网络会参与微调），或者全部使用（即我们使用 $\boldsymbol{z}_i$ 来进行预训练和模型微调）。

对于 N 个已观测到的训练信息级联图，训练损失函数被定义为均方对数误差（Mean Logarithmic Squared Error, MSLE）：

$$\mathcal{L}^{\text{supervised}} = \frac{1}{N}\sum_{i=1}^{N}\left(\log \hat{P}_i(t_p) - \log P_i(t_p)\right)^2, \tag{4-8}$$

其中 $P_i(t_p), \hat{P}_i(t_p)$ 分别是真实规模和预测的规模。

4.6.2 半监督学习和模型蒸馏

与之前的工作类似[124,134]，我们将监督学习的损失函数和无监督对比损失函数结合起来作为最终的目标函数：

$$\mathcal{L}^{\text{semi}} = \mathcal{L}^{\text{supervised}} + \theta_{\text{loss}}\mathcal{L}^{\text{contrastive}}, \tag{4-9}$$

其中 θ_{loss} 是一个权重超参数来平衡两个损失函数，$\mathcal{L}^{\text{contrastive}}$ 由所有的正样本来计算，当做模型的一个正则项，它强迫模型在对有标签信息级联图做出规模预测的时候，同时对无标签信息级联图和有标签信息级联图的增强视角进行判别。但是，这个损失函数的设计可能会遇到“负面迁移”问题[116]。受到 [125, 134] 启发，我们采用了两个独立的网络来缓解这个问题：一个微调过后的网络作为教师，另外一个初始化的网络作为学生。我们强迫学生网络的预测尽可能地与教师网络相似，

训练过程的损失函数定义如下：

$$\mathcal{L}^{\text{semi}} = \frac{1}{N+U}\sum_{i=1}^{N+U}\left(\log \hat{P}_i^T(t_p) - \log \hat{P}_i^S(t_p)\right)^2, \tag{4-10}$$

其中 N 是有标签数据中样本的数量，U 是无标签数据中样本的数量，$\hat{P}_i^T(t_p)$ 和 $\hat{P}_i^S(t_p)$ 分别是教师网络和学生网络的预测。在这种设置下，CCGL 可以同时受益于有标签数据和无标签数据：教师网络输出伪标签（Pseudo label）给学生网络。学生网络的结构可以和教师网络一致（自蒸馏，Self-distillation），或者是一个小的蒸馏网络。

4.7 与互信息最大化的联系

在理论中，学习信息级联图表示的过程可以被看做是最大化两个增强后的信息级联图视角 $\hat{\mathcal{G}}_i^1$ 和 $\hat{\mathcal{G}}_i^2$ 之间的互信息（Mutual information）[121,123]：

$$\mathcal{I}_{\text{MI}}(\Phi(\hat{\mathcal{G}}_i^1); \Phi(\hat{\mathcal{G}}_i^2)), \tag{4-11}$$

这两个信息级联图视角应该属于相同的类别或者在嵌入空间中的位置相近。其中 $\Phi(\cdot)$ 是一个神经网络，该网络中的图编码器将信息级联图 $\mathcal{G}_i$ 的 $\boldsymbol{h}_i$ 用作下游任务，而基于多层感知机的映射网络将 $\boldsymbol{h}_i$ 映射为 $\boldsymbol{z}_i$ 用作对比学习。从概率的角度看，给定两个随机变量 $\boldsymbol{z}_i^1$ 和 $\boldsymbol{z}_i^2$，模型被强迫去从来自于联合分布 $p(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2)$ 的正样本对和来自于 $p(\boldsymbol{z}_i^1)p(\boldsymbol{z}_i^2)$ 的负样本对之间做出判别。最小化对比损失函数等价于去最大化互信息 $\mathcal{I}_{\text{MI}}(\boldsymbol{z}_i^1; \boldsymbol{z}_i^2)$ 的一个下界，即：

$$\mathcal{I}_{\text{MI}}(\boldsymbol{z}_i^1; \boldsymbol{z}_i^2) = \mathbb{E}_{p(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2)}\left[\log \frac{p(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2)}{p(\boldsymbol{z}_i^1)p(\boldsymbol{z}_i^2)}\right] \tag{4-12}$$

$$\geq \log(2B) - \mathbb{E}\left[\mathcal{L}_{1,2}^{\text{contrastive}}(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2, 2B)\right]. \tag{4-13}$$

对比学习框架的核心部分在于数据视角的设计。在 CCGL 模型中，我们提出了基于模拟网络中信息传播过程的AugSIM数据增强策略来对信息级联图进行增强。AugSIM策略在捕获信息传播中的变化和不确定性的时候，还可以保留图视角之间的高阶共享信息，也就是说，学习到的信息级联图表示应该可以抵抗来自外界和不确定性的随机扰动。一方面，模型试着去对增强的正负样本图视角进行判别（最大化互信息），另一方面，我们对模型进行优化来使其忽略图视角之间不重要的差别，从而只保留重要的信息（最小化下游任务的预测误差）。

表 4-2 CCGL 模型在 58,489 个微博转发信息级联图上的运行时间分析

阶段	监督学习	CCGL
图数据增强	-	71.5s (1.2ms per)
预处理	17.5m (27ms per)	53.5m (55ms per)
预训练	-	14.0m (14ms per)
训练/微调	23.8m (36ms per)	14.9m (23ms per)
模型知识蒸馏	-	36.5m (28ms per)
总的运行时间	41.3m (63ms per)	107.4m (110ms per)

4.8 计算复杂度

与监督模型相比，CCGL 模型有三个模块带来了额外的计算消耗：（1）图数据增强模块；（2）对比自监督预训练模块；（3）模型知识蒸馏模块。

4.8.1 图数据增强模块

令 $|\mathcal{V}|$ 和 $|\mathcal{V}^{\text{leaf}}|$ 代表信息级联图中节点和（在添加节点的过程之后的）叶子节点的数量，数据增强策略AugSIM需要对图中的每一个节点进行遍历来执行增加节点的过程，然后对图中的每一个叶子节点进行遍历来执行删除节点的过程。AugSIM所需要的时间复杂度为 $\mathcal{O}(|\mathcal{V}| + |\mathcal{V}^{\text{leaf}}|)$，近似于或少于 $\mathcal{O}(2|\mathcal{V}|)$，其时间复杂度与图中节点的数量成线性关系。对于大约 58,000 个微博转发信息级联图，AugSIM只需要运行大约 71.5 秒。

4.8.2 预训练、模型微调和知识蒸馏

在预训练网络中，可训练的模型参数的数量（708K）与监督学习的基准模型（686K）相差不大。额外的参数主要来自于映射网络（增加的参数数量取决于映射网络的深度和广度）。在每一轮的预训练之中，CCGL 模型大约花费 27 秒；在每一轮的模型微调之中，CCGL 模型大约花费 13 秒；在每一轮的知识蒸馏之中，CCGL 模型大约花费 33 秒。

本章中的实验是在以下平台上运行的：系统为 Ubuntu 20.04，内存为 48GB，CPU 为 Intel® Core™ i7-8700K，GPU 为单个 NVIDIA 1080Ti。CCGL 模型由 Python 3.7、TensorFlow 2.3、CUDA 10.1 等实现。与监督模型的运行时间分析对比请见表4-2。我们在本节中所做的实验的默认参数请见表4-4。

表 4-3 五个信息级联数据集的统计信息

数据集	微博	推特	ACM	APS	DBLP
有标签信息级联的数量	39,076	18,198	12,988	27,802	4,879
无标签信息级联的数量	19,413	7,396	20,013	44,921	264
平均节点数量	174.02	141.58	17.69	30.82	14.99
平均观测到的节点数量	98.91	82.46	12.20	17.54	10.80
平均路径长度	2.2462	2.2111	2.0805	2.9164	2.5413

表 4-4 超参数设定

参数	搜索空间	值
数据增强力度 η	$\{0.05, 0.1, 0.2, 0.5, 1.0\}$	0.1
数据增强策略	`AugSIM`, `AugRWR`	`AugSIM`
Batch size B	$\{16, 32, 64, 256\}$	64
早停耐心	-	20
嵌入维度	$\{16, 32, 64, 128, 256, 512\}$	64
学习率	-	$5e^{-4}$
损失函数权重 θ_{loss}	-	1.0
预训练轮次	$\{10, 20, 30, 50, 100, 200, 500\}$	30
映射网络（100%）	从 0 到 4 层	4-1
映射网络（10%）	从 0 到 4 层	4-4
映射网络（1%）	从 0 到 4 层	4-3
重启概率 s	-	0.2
模型大小	$\{1\times, 2\times, 4\times, 8\times, 16\times\}$	4×
基于重启的随机游走步长 γ	$\{1.0, 2.0, 3.0, 5.0\}$	3.0
温度参数 τ	$\{0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$	0.1

4.9 实验

遵循常见的无监督和半监督学习研究[118,125,130]的实验实践, 我们分别在节4.9.1和节 4.9.2中总结了实验设置和多个信息级联数据集，在节4.9.3中介绍了基准模型及其实验配置. 在节4.9.4中讨论了实验结果、多个实验发现和消融实验。为了评估 CCGL 模型的泛化能力, 我们在节4.9.5中测试了 CCGL 模型在不同的信息级联预测任务上和不同的信息级联数据集上的知识迁移性能。

4.9.1 实验设置

对于 CCGL 模型和基准模型的所有实验，除非另有说明，为了公平比较起见，我们统一采用以下设置。我们使用 Adam 优化器，每个数据集划分为训练集（50%）、验证集（10%）和测试集（40%）以及未标记的信息级联数据。当训练

（或验证）损失连续 20 轮的训练中没有下降，预训练（或模型微调）过程将提前停止。根据之前的研究[21,65,78]，我们使用以 2 为底的均方对数误差（MSLE）作为损失函数。节点数在观察时间内小于 10 的信息级联将会被过滤掉，对于节点数在观察时间内大于 100 的信息级联图，我们只选择其前 100 个节点（按采用时间排序）。

我们通过搜索超参数空间来手动调整模型的超参数。表 4-4 列出了本文中使用的超参数、搜索空间及其默认值。

4.9.2 数据集

我们使用了 5 个公开的大规模信息级联数据集，这些数据集可分为两类：社交网络数据集和科研论文数据集。它们的详细统计请见表 4-3。

- **微博转发数据集**：来自于论文 [21]。对于该数据集中的信息级联，我们将其观察时间设置为 1 小时，预测时间设置为 24 小时。
- **推特标签数据集**：来自于论文 [87]。对于该数据集中的信息级联，我们将其观察时间设置为 2 天，预测时间设置为 32 天。推特数据集中的信息级联图是由推特之间的转发、参与、提醒等关系构建的。
- **ACM 引用数据集**：是由发表在美国计算机协会（Association for Computing Machinery, ACM）上的论文所构建的，该数据集公布于 2017 年 1 月 20 日，其中包含了 2,385,057 篇计算机科学领域的科研论文。我们将该数据集中信息级联的观察时间设置为 3 年，预测时间设置为 10 年。
- **APS 引用数据集**：是由发表在美国物理学会（American Physical Society, APS）上的论文所构建的，该数据集的获取时间为 2019 年 1 月 17 日，包含了 616,316 篇发表在 APS 期刊上的科研论文。我们将该数据集中信息级联的观察时间设置为 3 年，预测时间设置为 20 年。
- **DBLP 引用数据集**：来自于 DBLP 引文网络 v9[136]，发布于 2017 年 7 月 3 日，其中包含了 3,680,006 篇科研论文。我们将该数据集中信息级联的观察时间设置为 5 年，预测时间设置为 20 年。前五年内引用数量少于 5 次的论文将被筛选掉。

4.9.3 基准模型

为了评估 CCGL 模型的性能，以及展示其三个主要模块的优点，即不基于特定任务的对比自监督预训练来提升模型泛化能力、基于图的数据增强策略来捕捉信息传播中的变化和不确定性、基于特定下游任务的模型微调和模型知识蒸馏来进行知识迁移学习，我们选取了多个强大的监督和半监督模型，它们的介绍如下：

4.9.3.1 监督模型

- **基于特征工程的模型**：从信息级联中抽取各种有用的人工设计的特征来对信息级联的规模做出预测。依据之前的论文[12,78]，我们选取了如下结构特征和时间特征：累积规模序列、根节点发布时间和第一个参与者时间之间的时间差，所有参与者们的前半部分的平均参与时间，所有参与者们的后半部分的平均参与时间，叶子结点的数量、平均节点度、平均和最大传播路径长度等。抽取到的信息级联特征作为多层感知机的输入，然后进行训练和预测。
- **node2vec**[72]：是一个基于随机游走的节点嵌入模型。我们使用它来获取信息级联图中节点的嵌入，并将获取到的节点嵌入输入到双向门控循环单元和多层感知机中来做出预测。我们将节点嵌入的维度设置为 128，窗口大小设置为 10，步长设置为 30，步的数量设置为 200，node2vec 模型中对邻居进行采样的参数 p 和 q 分别设置为 1。node2vec 模型的实现来自于https://github.com/eliorc/node2vec。
- **DeepHawkes**[21]：结合了霍克斯自激励过程和深度学习技术来对信息级联的规模做出预测。DeepHawkes 模型中有三个关键的霍克斯过程模块，即用户影响力、自激励机制、以及时间衰减因素。它们在实现中分别由用户嵌入、基于和池化（Sum pooling）的路径编码、以及非参数化的时间衰减函数来建模。DeepHawkes 模型的嵌入维度是 64，学习率是 $5e^{-3}$，嵌入学习率是 $5e^{-4}$，L_2 系数是 $5e^{-2}$，dropout 概率是 0.8，时间间隔是 5 分钟。DeepHawkes 模型的实现来自于https://github.com/CaoQi92/DeepHawkes。
- **基础模型（Base）**[78]：在本章中是一个标准的监督模型，它与微调的 CCGL 模型拥有相同的网络结构（例如信息级联图编码器和多层感知机等）。与 CCGL 模型相比，它缺少对比自监督训练和模型知识蒸馏。为了公平对比起见，Base 模型的超参数设置和 CCGL 模型的超参数设置完全相同。

4.9.3.2 半监督模型

- **变分自编码器**（Variational Auto-Encoder）[102] 是一类深度生成式模型，它基于非监督表示学习。它的编码器和解码器均基于门控循环单元和多层感知机。信息级联中节点嵌入的序列作为模型的输入来生成隐藏表示 $\boldsymbol{z}$，然后将该表示输入到解码器中来重构模型的输入（即节点的嵌入）。我们将变分自编码器中的变分模块移除，从而创建一个新自编码器基准模型（AE）。自编码器模型由重构损失来优化，变分自编码器的优化过程还包括一个 ELBO 证据下界项。我们将这两个模型的预训练轮次设置为 100.

4.9.3.3 另一种数据增强策略

- **AugRWR**：为了展示我们在 CCGL 模型中设计的 AugSIM 数据增强策略的有效性，部分受 GCC[130] 模型启发，我们实现了一种新的数据增强策略 AugRWR，它基于带有重启的随机游走（RWR）来对信息级联图进行数据增强。通过重复带有重启的随机游走步骤，我们对信息级联图中的节点集合 $\mathcal{V}_i$ 采样到一系列子节点集合，我们将其称为 $\mathcal{V}_i^{\text{rwr}}$。然后增强后的信息级联图可以表示为 $\tilde{\mathcal{G}}_i$，它是从原信息级联图的节点集合中移除以下节点 $\mathcal{V}_i \setminus \mathcal{V}_i^{\text{rwr}}$ 来构成的。我们限制游走过程最多 $\gamma|\mathcal{V}_i|$ 次，来避免对小规模的信息级联图采样到完全相同的视角，以及避免对大规模的信息级联图采样不完全。带有重启的随机游走的转移概率（Transition probability）是由信息级联图中邻居节点的度来确定的：

$$p(v_j^i \in \mathcal{N}(u_j^i)|u_j^i) = \frac{\text{degree}(v_j^i)}{\sum_{v_k^i \in \mathcal{N}(u_j^i)} \text{degree}(v_k^i)}, \tag{4-14}$$

其中 $\mathcal{N}(u_j^i)$ 是节点 u_j^i 的邻居节点集合。

4.9.4 实验结果与分析

CCGL 模型与基准模型的实验对比结果请见表4-5和表4-6。其中包括三类模型：监督模型、在固定特征上的线性评估模型、以及半监督微调模型。实验结果表明，与其他基准模型相比，我们提出的 CCGL 模型显著地提升了预测效果。我们从以下三个角度给出具体的讨论：（1）回答节4.2提出的三个研究问题；（2）两个重要的观察；以及（3）四个消融实验。

4.9.4.1 当对模型进行知识蒸馏之后，使用无标签数据可以提升预测的效果

表4-6的结果表明，在使用大规模无标签数据且没有模型蒸馏的时候，CCGL 的预测效果有一些轻微下降，尤其是当有标签数据稀少、模型被线性评估的情况下。我们猜测效果下降的原因在于，无标签数据的引入影响了信息级联的表示学习，扰乱了其嵌入空间。当我们使用了模型知识蒸馏之后，效果得到了显著地提升。引入知识蒸馏的另一好处在于模型的泛化能力得到了提升。更多有关模型迁移学习的内容请见节4.9.5。

4.9.4.2 我们设计的信息级联数据增强策略在对比学习中是有效的

信息级联数据增强策略AugSIM和AugRWR均提升了模型的预测效果。当有标签数据稀少的情况下，提升的比例更大。AugSIM策略的效果要比AugRWR策略好，这

表 4-5 模型的设置：表中第一列的数字为模型的编号，与表4-6中的编号对应。

#	模型	使用无标签	数据增强策略	损失函数
	监督学习基准模型：			
1	Feature	-	-	监督
2	DeepHawkes [21]	-	-	监督
3	node2vec [72]+BiGRU	-	-	监督
4	Base (rand. init.) [78]	-	AugSIM	监督
5		-	-	监督
	半监督模型（包括完全微调的和线性评估的）：			
6	AE	✓	-	重构损失
7	VAE	✓	-	重构损失 +ELBO
8	CCGL（固定参数）	✓	AugRWR	监督 + 对比
9		✗	AugSIM	监督 + 对比
10		✓	AugSIM	监督 + 对比
11	CCGL	✓	AugSIM	监督 + 对比
12		✓	AugRWR	监督 + 对比
13		✗	AugSIM	监督 + 对比
14		✗	AugRWR	监督 + 对比
15		✓	AugSIM+AugRWR	监督 + 对比
16		✓	AugRWR	监督 + 无监督 + 对比
17		✓	AugSIM	监督 + 无监督 + 对比

一点证明了相比基于随机游走的策略，我们设计的模拟网络中信息传播的数据增强方法是有效的。特别地，在 1%、10%、100% 的有标签数据上进行模型微调的时候，相比于 Base 模型，CCGL 模型大约提升了 5.3%、9.3%、1.4% 的预测效果。这进一步证明，在信息级联预测中考虑信息传播的变化和不确定性有益于模型预测和减轻过拟合现象。

4.9.4.3 CCGL 模型的预测效果高于强大的监督模型

完整的 CCGL 模型，即使用无标签信息级联图数据、使用为信息级联图特别设计的数据增强方法AugSIM、使用图对比自监督学习进行预训练、使用基于特定任务的模型微调和知识蒸馏，CCGL 模型取得了非常好的信息级联规模预测效果。与强大的有监督模型 Base 相比，CCGL 模型在微博数据集上使用 1%、10%、100% 的标签数据的时候，分别取得了 9.2%、11.7%、2.9% 的预测效果提升。

此外，我们还有以下两个重要的观察。

表 4-6 CCGL 模型与基准模型在不同实验条件下的实验结果对比。表中第一列的数字于实验设置表4-5中的数字对应。评价指标为运行十次的平均均方对数误差（MSLE），数值越低效果越好。我们使用了标签数据数量不同的微博数据集。其他四个数据集（推特、ACM、APS、DBLP）的实验结果对比请见表4-13。我们使用了自蒸馏的学生网络。均方对数误差的右下角的数字代表了标准差，右上角的橙色数字代表了相比于 Base 模型预测效果的提升值（至少 0.05）。

#	模型	微博 (1%)	微博 (10%)	微博 (100%)
	监督学习基准模型：			
1	Feature	$4.94_{\pm 0.70}$	$3.35_{\pm 0.15}$	$3.01_{\pm 0.17}$
2	DeepHawkes [21]	$4.39_{\pm 0.10}$	$3.41_{\pm 0.03}$	$3.24_{\pm 0.02}$
3	node2vec [72]+BiGRU	$3.76_{\pm 0.38}$	$3.44_{\pm 0.03}$	$2.95_{\pm 0.03}$
4	Base (rand. init.) [78]	$3.82_{\pm 0.10}$	$3.34_{\pm 0.01}$	$3.06_{\pm 0.19}$
5		$\mathbf{3.58}_{\pm 0.03}$	$\mathbf{3.24}_{\pm 0.09}$	$\mathbf{2.77}_{\pm 0.04}$
	半监督模型（包括完全微调的和线性评估的）：			
6	AE	$4.00_{\pm 0.38}$	$3.66_{\pm 0.57}$	$2.78_{\pm 0.04}$
7	VAE	$3.96_{\pm 0.44}$	$3.63_{\pm 0.71}$	$2.78_{\pm 0.09}$
8	CCGL（固定参数）	$4.18_{\pm 0.05}$	$3.81_{\pm 0.01}$	$2.84_{\pm 0.02}$
9		$4.42_{\pm 0.23}$	$3.69_{\pm 0.02}$	$2.81_{\pm 0.04}$
10		$4.51_{\pm 0.09}$	$4.11_{\pm 0.04}$	$2.78_{\pm 0.02}$
11	CCGL	$3.39^{+0.19}_{\pm 0.10}$	$2.94^{+0.30}_{\pm 0.03}$	$2.73_{\pm 0.02}$
12		$3.49^{+0.09}_{\pm 0.05}$	$3.12^{+0.12}_{\pm 0.29}$	$2.74_{\pm 0.02}$
13		$3.38^{+0.20}_{\pm 0.04}$	$2.95^{+0.29}_{\pm 0.04}$	$2.73_{\pm 0.02}$
14		$3.44^{+0.14}_{\pm 0.05}$	$3.04^{+0.20}_{\pm 0.08}$	$2.75_{\pm 0.04}$
15		$3.35^{+0.23}_{\pm 0.08}$	$2.96^{+0.28}_{\pm 0.04}$	$2.71^{+0.06}_{\pm 0.02}$
16		$3.40^{+0.18}_{\pm 0.01}$	$3.02^{+0.22}_{\pm 0.01}$	$2.72^{+0.05}_{\pm 0.01}$
17		$\mathbf{3.25}^{+0.33}_{\pm 0.02}$	$\mathbf{2.86}^{+0.38}_{\pm 0.01}$	$\mathbf{2.69}^{+0.08}_{\pm 0.00}$

4.9.4.4 观察一：相比于基准模型，CCGL 模型需要更少的标签数据

当在不同比例的有标签数据上进行微调的时候，CCGL 模型比其他基准模型更少地依赖于有标签数据。当只有 1% 的有标签数据可用的时候，CCGL 模型的效果与在 10% 有标签数据上训练的 Base 模型不相上下（3.25 vs. 3.24）。这一提升归功于 CCGL 模型中的数据增强模块和对比自监督学习。

4.9.4.5 观察二：有监督学习模型并不能从数据增强中受益

当我们使用AugSIM来增强原始的信息级联图的时候，我们发现这并不能提升有监督学习模型的预测效果。事实上，引入AugSIM数据增强策略显著地降低了它们的预测效果。当标签数据越来越多的时候，模型预测效果下降的越多。一种可

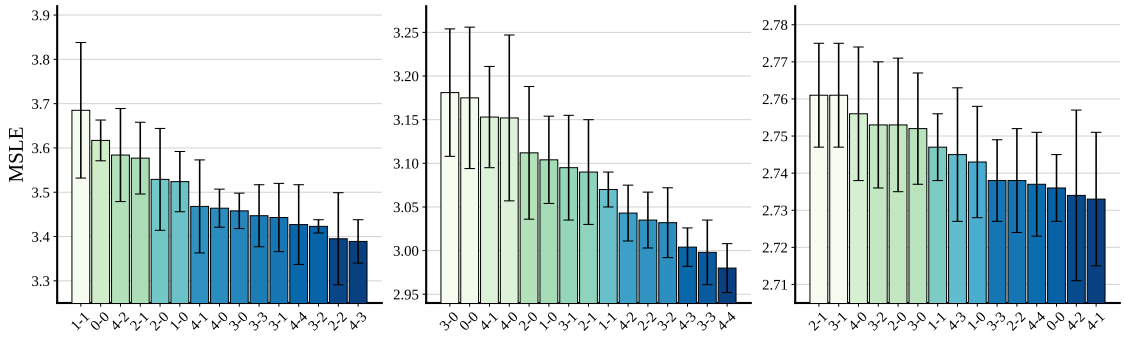


图 4-4 映射网络的不同设计对预测的影响。我们使用不同标签比例的微博数据集。横轴的标签 *i-j* 表示映射网络的深度是 *i*，我们从映射网络的第 *j* 层来微调 CCGL 模型。我们运行五次实验来汇报平均均方对数误差（MSLE）及其标准差。左图：在 1% 的有标签数据上微调；中间图：在 10% 的有标签数据上微调；右图：在 100% 的有标签数据上微调。

能的原因解释是：有监督模型学习特征级别的表示，而不是抽象级别的语义表示，这导致它们没有办法捕获到数据增强所带来的变化和不确定性。

为了更好地显示 CCGL 模型的鲁棒性，我们还做了以下四个消融实验。

4.9.4.6 消融实验一：深的映射网络对信息级联自监督学习更有效

我们发现，深层的映射网络对信息级联自监督学习更为有效，尤其是在有标签数据很少的时候。之前的工作 [124] 发现，引入一个非线性的、可训练的、基于多层感知机的映射网络（Projection head）可以显著地提升视觉表示学习的效果。随后的工作 [135] 也证明了这一点。在 [125] 中，作者们发现一个深层的映射网络，以及从映射网络的中间层开始进行模型微调，可以提升图片分类的效果。但是，深层的映射网络会不会提升信息级联预测任务的效果还没有人研究过。我们尝试了 15 种不同的映射网络设计，实验结果见图4-4。我们将映射网络的设计标记为 *i-j*，其中 *i* 代表映射网络总的深度，它的值从 0 到 4，*j* 表示 CCGL 模型从映射网络的第 *j* 层进行模型微调。在有标签数据的数量不同的情况下，映射网络的效果也不同。当在 1% 或者 10% 的有标签数据上进行微调的时候，映射网络显著地提升了预测效果，在 14 种映射网络的设计中，有 13 种设计超过了 0-0 的预测效果（即没有映射网络），在 1% 的标签上效果提升最高达到 6.3%，在 10% 的标签上效果提升最高达到 6.2%。当有标签数据变多的时候（即 100%），14 中设计中只有 2 种设计提升了预测效果。这一现象提供了一种使用映射网络反而带来负面作用的例子。对于映射网络的深度，我们发现深的映射网络的预测效果要好于浅的映射网络。如表4-7所示，与之前工作 [125] 的结论不同，我们没有发现从中间的映射层来

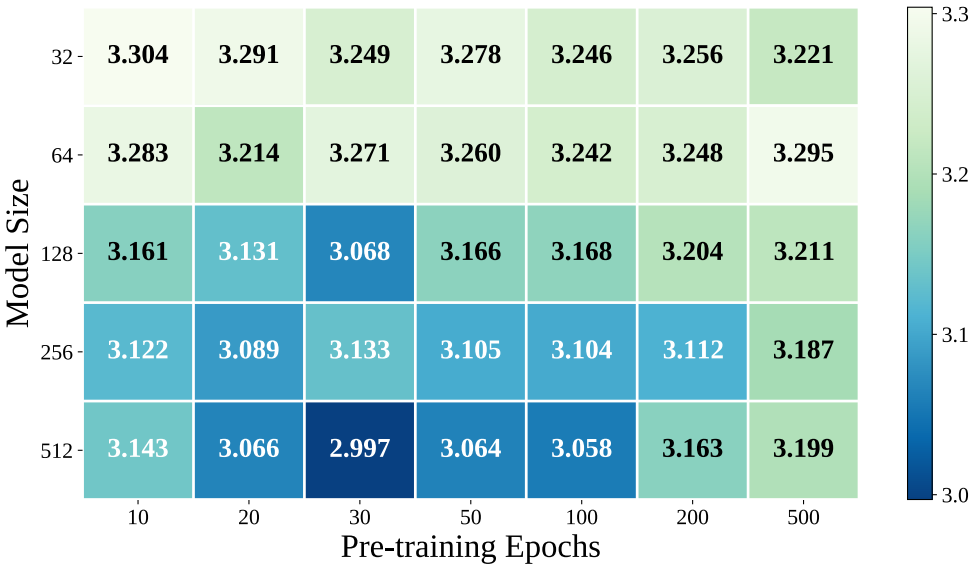


图 4-5 不同模型大小和预训练轮次的组合对模型预测的影响。我们使用具有 10% 标签数据的微博数据集。我们运行五次实验来汇报平均均方对数误差（MSLE）。

微调网络是更好的选择。

表 4-7 从映射网络的中间层进行微调并不总是更好的选择。

标签比例	第一层	中间的层	最后一层
1%	3.484	**3.470**	3.479
10%	3.127	3.060	**3.011**
100%	2.741	2.736	**2.730**

映射网络在视觉表示学习中带来了普遍地效果提升，但是对于信息级联学习来说，要不要使用映射网络，或者使用哪种映射网络，还没有一个统一的结论或指导。

4.9.4.7 消融实验二：模型大小和预训练轮数

在图4-5中，我们展示了不同模型大小和预训练轮数组合对模型效果所带来的影响。在这里我们将基础模型大小表示为 1×，即模型的宽度（包括嵌入维度、循环神经网络的单元数、多层感知机的单元数等）设置为 32。其他超参数的设置为：batch 大小为 64，数据增强策略为`AugSIM`，数据增强力度 η 为 0.1，温度参数为 0.05，映射网络设计为 2-0。从结果中我们可以发现，大的模型可以保证一个比较满意的预测效果。当模型的大小已经很大的时候，增加预训练轮数带来的提升很小，甚至有时候会降低模型的预测效果，这种情况可能由负面预训练而导致的。

最佳的预测效果由大小为 16× 的模型预训练 30 轮次而得到的。

表 4-8 模型大小对监督学习和半监督学习的影响。我们使用具有 10% 标签数据的微博数据集。我们运行五次实验来汇报平均均方对数误差（MSLE）。

模型	**1×**	**2×**	**4×**	**8×**	**16×**
监督模型	$3.32_{\pm 0.02}$	$3.29_{\pm 0.11}$	$3.21_{\pm 0.15}$	$3.16_{\pm 0.16}$	$3.30_{\pm 0.06}$
半监督模型	$3.22^{+0.10}_{\pm 0.06}$	$3.21^{+0.08}_{\pm 0.09}$	$3.07^{+0.14}_{\pm 0.04}$	$3.09^{+0.07}_{\pm 0.04}$	$2.99^{+0.31}_{\pm 0.02}$

我们还研究了模型大小对监督学习和半监督学习模型的影响。实验结果请见表4-8。从结果中我们知道，对于所有的模型大小，半监督学习模型都超过了有监督学习模型。当模型大小变大的时候，有监督学习模型容易陷入到过拟合，而半监督学习模型很大程度上缓解了这一点，这可能是由于以下两个原因：（i）结合有标签数据和无标签数据可以使模型的泛化能力提高；（ii）监督学习模型依赖于细粒度的特征级别的表示学习，而对比学习模型学习到了高阶的抽象语义表示。

4.9.4.8 消融实验三：知识蒸馏可以带来额外的效果提升

表 4-9 不同的对教师网络和学生网络的知识蒸馏实验设定。我们使用了带有不同比例标签数据的微博数据集。我们汇报均方对数误差（MSLE）。

教师	**学生**	**1%**	**10%**	**100%**
Base	-	3.58	3.24	2.77
微调网络	没有知识蒸馏	3.39	2.94	2.73
有标签	有标签，自蒸馏	3.28	2.88	2.70
有标签 + 无标签	有标签，自蒸馏	3.27	2.87	2.71
有标签 + 无标签	无标签，自蒸馏	3.29	2.89	2.84
有标签 + 无标签	有标签 + 无标签，蒸馏	3.26	2.88	2.77
有标签 + 无标签	有标签 + 无标签，自蒸馏	**3.25**	**2.86**	**2.69**

我们发现，与微调的模型相比，同时在无标签数据和有标签数据上进行模型的知识蒸馏可以带来额外的效果提升。我们使用了两种不同的教师网络和四种不同的学生网络来进行模型知识蒸馏。实验的结果请见表4-9。教师网络是在有标签数据上进行预训练，或者同时在有标签数据和无标签数据上进行预训练。而学生网络是在（1）有标签数据；（2）无标签数据；（3）同时在有标签和无标签数据上进行知识蒸馏。从结果中我们发现，模型蒸馏确实可以带来额外的效果提升，与对比学习带来的提升相比，在使用 1% 的有标签数据的时候，知识蒸馏最高可以多

带来 73.7% 的相对提升。在使用 10% 的有标签数据的时候，知识蒸馏最高可以多带来 26.7% 的相对提升。在使用完整的有标签数据的时候，知识蒸馏最高可以带来 50% 的相对提升。知识蒸馏起作用的一种可能的解释是：蒸馏可以是模型变得更加的不基于特定任务，从而缓解负面迁移的问题。

4.9.4.9 消融实验四：超参数分析

我们现在分析 CCGL 模型中重要的超参数对模型效果的影响。我们使用带有不同比例标签数据的微博数据集。实验结果请见图4-6。

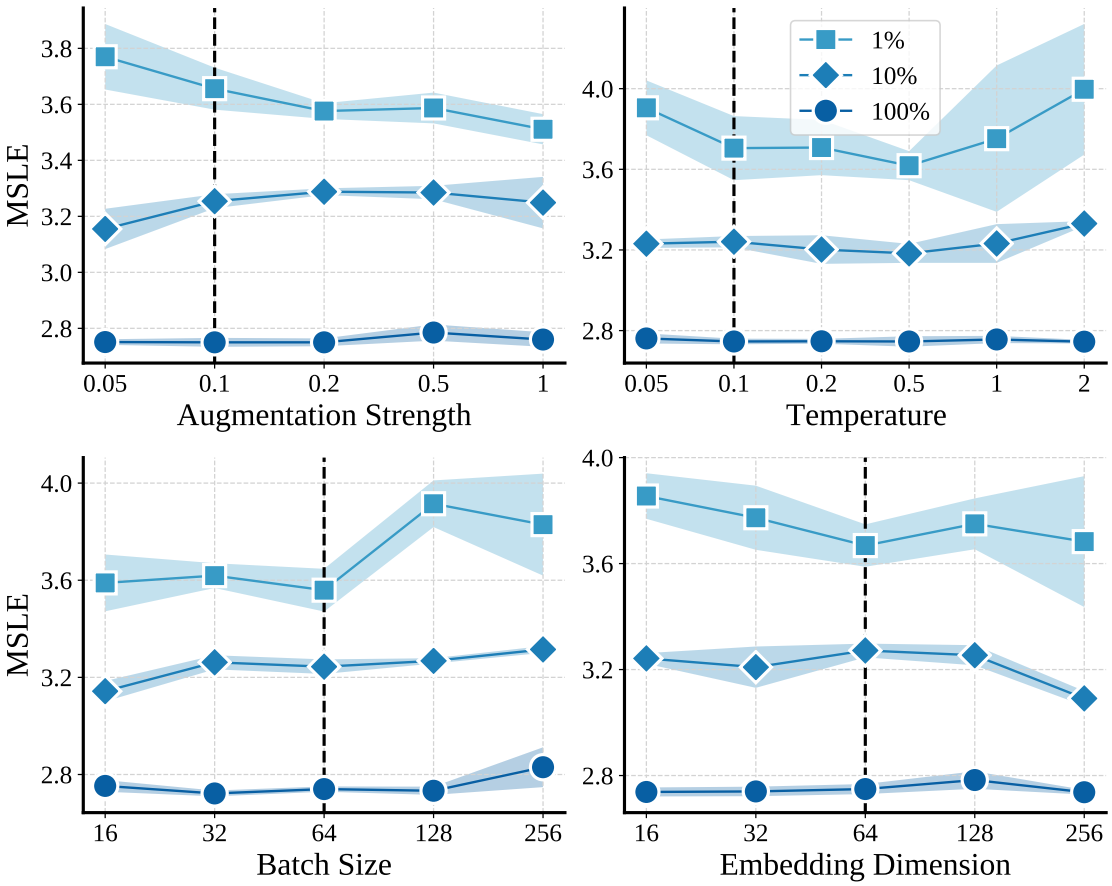


图 4-6 在带有不同比例标签数据的微博数据集上对 CCGL 中的重要超参数进行实验分析。垂直线代表本章实验中所使用的默认设置。我们运行五次实验来汇报平均均方对数误差（MSLE）及其标准差。

我们默认使用了下面的超参数设置：batch 大小为 64，数据增强策略为`AugSIM`，数据增强力度 η 为 0.1，温度参数 τ 为 0.05，预训练轮数为 100，嵌入维度为 64，模型大小为 64（2×），映射网络为两层全连接层，我们在映射网络之前对模型进行微调，即 2-0。

超参数一 **数据增强力度 η 的影响**：数据增强力度 η 控制着信息级联图中增加的或者删除的节点及边的数量，我们发现在 1% 的有标签数据上进行预训练的时候，大的数据增强力度有益于模型预测。

超参数二 **对比损失函数中的温度参数 τ 的影响**：实验结果表明，一个小的温度参数（大约在 0.2 左右）是比较合适的。大的温度参数（例如 1 或 2）有时候会让模型难以收敛（我们进行了 30 次实验，有 8 次实验没有收敛）。

超参数三 **batch 大小 B 的影响**：与之前工作 [124] 的结论不同，信息级联图表示学习可能更偏好于小的 batch。当数据集中所有的有标签数据被使用时，大的 batch（例如 128 或 256）会让模型的预测变差以及变得不稳定。当只有 1% 的有标签数据可用的时候，这一情况变得更加严重。这一结果表明，大的 batch 在对比学习总并不是必须的。我们猜测这是由于在信息级联图建模和预测之中，数据的变化和信息要显著地少于视觉数据（例如 ImageNet 数据集），所以小的 batch 已经足够使模型去分辨不同的信息级联。类似的结论在 [130] 中也有提及，构建了一个大的 batch（也被称为字典大小）并不总对模型的预测有帮助，有些时候甚至会降低模型的预测效果。不过，如果信息级联中蕴含了足够多的信息（例如可以包含信息中文字、图像、全局图等），一个非常大的 batch（例如 4096 或 8192），也许会显著地提升模型的预测效果。我们将这个猜想留作未来的工作。

超参数四 **嵌入维度的影响**：我们更改嵌入的维度来探究其对模型预测的影响，维度值从 16 到 256。结果显示，大的嵌入维度对提升预测的效果有帮助。但是大的嵌入维度也会带来额外的时间和空间消耗，而且在有标签数据很少的时候，容易使模型陷入过拟合。

4.9.5 知识迁移学习

为了研究我们提出的 CCGL 模型的泛化性能，我们在五个信息级联数据集（微博、推特、ACM、APS、DBLP）以及两个信息级联预测任务（规模预测、爆发预测）上对 CCGL 模型的迁移能力做了实验。

4.9.5.1 在不同的信息级联数据集上进行知识迁移

为了展示 CCGL 模型学习到的信息级联表示具有在不同信息级联数据集间进行知识迁移的能力，我们在微博和推特数据集上做了如下实验：（1）在单个数

据集上进行自监督预训练，然后在另一个不同的数据集上对模型进行微调监督训练；（2）在两个数据集上同时进行自监督预训练，然后在一个数据集上对模型进行微调监督训练。实验的结果请见表4-10。

表 4-10 在不同的信息级联数据集上进行知识迁移。我们运行十次实验来汇报平均均方对数误差（MSLE）及其标准差。

预训练	模型微调	1%	10%	100%
模型随机初始化（微博）	-	$3.58_{\pm 0.03}$	$3.24_{\pm 0.09}$	$2.77_{\pm 0.04}$
模型随机初始化（推特）	-	$8.32_{\pm 0.28}$	$6.31_{\pm 0.15}$	$5.32_{\pm 0.05}$
微博	微博	$3.39^{+0.19}_{\pm 0.10}$	$2.94^{+0.30}_{\pm 0.03}$	$2.73_{\pm 0.02}$
推特	推特	$7.49^{+0.83}_{\pm 0.23}$	$5.81^{+0.50}_{\pm 0.06}$	$5.20^{+0.12}_{\pm 0.05}$
微博	推特	$7.38^{+0.94}_{\pm 0.09}$	$5.77^{+0.54}_{\pm 0.06}$	$5.23^{+0.09}_{\pm 0.05}$
推特	微博	$3.47_{\pm 0.11}$	$3.06^{+0.18}_{\pm 0.06}$	$2.75_{\pm 0.02}$
微博 & 推特	微博	$3.39^{+0.19}_{\pm 0.08}$	$2.98^{+0.26}_{\pm 0.02}$	$\mathbf{2.70}_{\pm 0.02}$
	推特	$\mathbf{6.89}^{+1.43}_{\pm 0.03}$	$\mathbf{5.70}^{+0.61}_{\pm 0.08}$	$\mathbf{5.14}^{+0.18}_{\pm 0.04}$

我们有两个重要的发现：（1）在微博数据集上进行预训练和在推特数据集上进行微调的 CCGL 模型的预测效果显著地超过了随机初始化的 Base 模型。当有标签数据很少的时候，它的预测效果甚至超过了同时在推特数据集上进行预训练和微调的模型。这个结果显示，微博数据集不仅帮助了在推特数据集上的信息级联预测，还提供了比推特数据集本身更好的模型微调起始参数；（2）当同时在微博和推特数据集上进行预训练的时候，微调的 CCGL 模型取得了比其他组合更好的预测效果，在只有 1% 的有标签数据可用的推特数据集上，相比于同时在推特数据集上进行预训练和微调的模型，它有着高达 72.3% 的相对提升（1.43 vs. 0.83）。CCGL 模型在预训练阶段学习到了通用的知识，在基于特定任务的微调阶段学习到了基于特定任务的知识。

4.9.5.2 将知识迁移到另外一个信息级联预测任务

除了我们之前研究的信息级联规模预测（Information cascade popularity prediction）任务，我们探究 CCGL 模型在另外一个信息级联下游预测任务上的知识迁移能力，即信息级联爆发预测（Information cascade outbreak prediction）。

对于五个信息级联数据集中的每一个，我们选取规模最大的前 10% 的信息级联作为爆发的信息级联，剩下的信息级联作为普通或者非爆发的信息级联。因为信息级联规模的分布是极度扭曲的，我们对非爆发信息级联进行欠采样来生成一个标签平衡的数据集。知识迁移的构成描述如下。

CCGL 模型首先在微博数据集上进行预训练。然后固定所有的超参数（这一

表 4-11 将知识迁移到信息级联爆发预测任务上。我们使用了所有五个带有不同比例有标签数据的信息级联数据集。我们使用准确率（Accuracy）作为评价指标。CCGL 模型在微博数据集上预训练 30 个轮次，然后分别在五个数据集上进行模型微调。

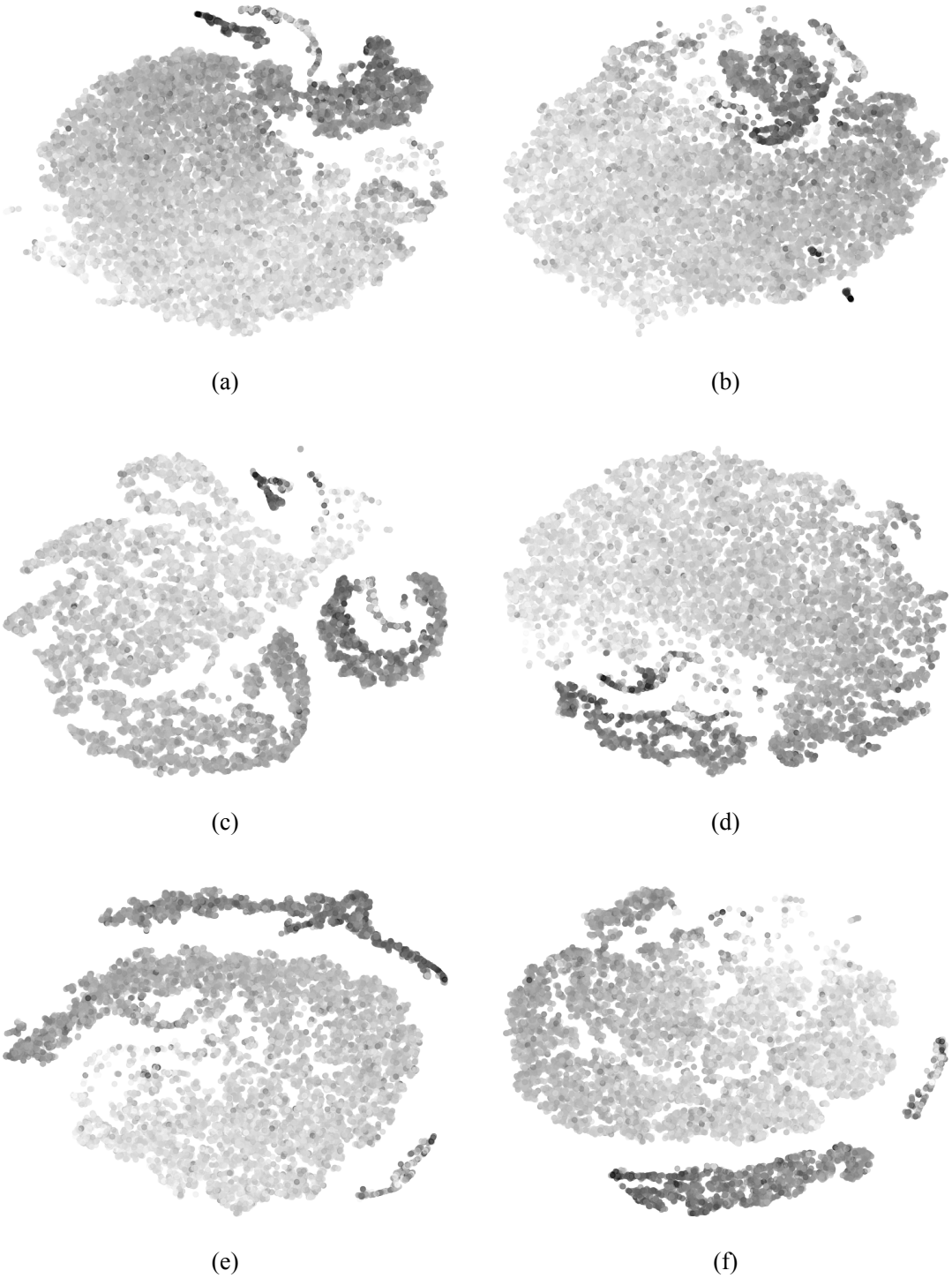
数据集	模型	1%	10%	100%
微博	随机初始化	82.64	82.37	83.70
	CCGL	81.89 (-0.75)	83.49 (+1.12)	83.90 (+0.20)
推特	随机初始化	61.50	78.44	84.06
	CCGL	81.48 (+19.9)	87.30 (+8.86)	87.12 (+3.06)
ACM	随机初始化	63.73	64.84	69.35
	CCGL	60.49 (-3.24)	68.86 (+4.02)	71.09 (+1.74)
APS	随机初始化	74.40	76.30	80.05
	CCGL	77.28 (+2.88)	79.94 (+3.64)	81.10 (+1.05)
DBLP	随机初始化	74.60	75.40	75.30
	CCGL	74.16 (-0.44)	76.96 (+1.56)	77.28 (+1.98)

点允许我们在进行知识迁移之前无需过多地对模型进行调参，并且保证一定的公平性，当然，这样做可能会降低预测的效果）。实验结果请见表4-11。总的来说，与随机初始化的 Base 模型相比，CCGL 模型在五个数据集上取得了有竞争力的预测效果。我们有如下两个重要观察：（1）当只有 1% 的有标签数据可用的时候，CCGL 模型极大地提升了在推特数据集上的预测效果，相比于 Base 模型，它提升了 19.9% 的预测准确率（61.50 vs. 81.48）；（2）尽管 CCGL 模型只在微博数据集上进行了预训练，它在其他的数据集上也表现的很好，包括社交网络数据集和科研论文数据集。这表示从微博数据集中预训练得来的知识被成功地迁移到了其他的数据领域。总的来说，在所有 15 个实验里，CCGL 模型在 12 个实验上取得了领先的效果。我们相信 CCGL 模型的通用知识学习能力和迁移能力应归功于在对比学习范式之下的无监督预训练，以及基于教师-学生框架的模型知识蒸馏。

4.9.6 其他实验

4.9.6.1 隐藏表示的可视化

在图4-7中，我们使用 *t*-SNE 对 CCGL 模型在微博数据集上学习到的信息级联表示进行可视化。图中的 (a) 和 (b) 分别是在有 19,538 条信息级联的微博训练集上预训练的 CCGL 模型中的表示 $\boldsymbol{h}$ 和 $\boldsymbol{z}$。在 CCGL 模型中，表示 $\boldsymbol{h}$ 被用作下游预测任务，而表示 $\boldsymbol{z}$ 用作计算对比损失函数。我们使用了结构为 4-1 的映射网络，嵌入维度是 256。图中的 (c) 到 (f) 是从不同模型中抽取的表示 $\boldsymbol{h}$。子图 (c) 是来自于有



(a) (b)

(c) (d)

(e) (f)

图 4-7 使用 *t*-SNE 来对学习到的信息级联隐藏表示进行可视化。(a) 和 (b) 分别是在有 19,538 条信息级联的微博训练集上预训练的 CCGL 模型中的 $\boldsymbol{h}$ 和 $\boldsymbol{z}$。(c) 到 (f) 分别是在有 15,630 条信息级联的微博测试集上使用 Base 模型，或者使用线性评估的、微调的、知识蒸馏的 CCGL 模型得到的 $\boldsymbol{h}$。

表 4-12 模型大小对监督学习模型和半监督学习模型预测效果的影响实验。我们运行十次实验并报告平均均方对数误差（MSLE）及其标准差。我们使用带有 1% 标签数据或 100% 标签数据的微博数据集。

模型	**1×**	**2×**	**4×**	**8×**	**16×**
在 1% 标签数据上进行训练：					
监督学习模型	$3.74_{\pm 0.24}$	$3.57_{\pm 0.08}$	$3.98_{\pm 0.35}$	$4.44_{\pm 0.07}$	$4.53_{\pm 0.04}$
半监督学习模型	$3.73_{\pm 0.04}$	$3.52^{+0.05}_{\pm 0.07}$	$3.47^{+0.52}_{\pm 0.03}$	$3.45^{+1.01}_{\pm 0.10}$	$3.51^{+1.02}_{\pm 0.05}$
在 100% 标签数据上进行训练：					
监督学习模型	$2.78_{\pm 0.05}$	$2.79_{\pm 0.04}$	$2.83_{\pm 0.03}$	$2.78_{\pm 0.02}$	$2.83_{\pm 0.03}$
半监督学习模型	$2.77_{\pm 0.05}$	$2.73^{+0.06}_{\pm 0.02}$	$2.75^{+0.08}_{\pm 0.01}$	$2.75_{\pm 0.01}$	$2.81_{\pm 0.05}$

监督训练的 Base 模型，子图 (d) 是来自于线性评估的 CCGL 模型，子图 (e) 是来自于微调的 CCGL 模型，子图 (f) 是来自于知识蒸馏的 CCGL 模型。以上模型使用 10% 的有标签数据训练。图中每个点（即一条信息级联）的值代表其未来的规模，点的颜色越深代表信息级联的规模越大。从图中可以看出，与预训练或线性评估的表示 $\boldsymbol{h}$ 相比，基于特定任务的微调或蒸馏模型 $\boldsymbol{h}$ 的节点聚类效果更为明显，这说明它们学习到的表示对信息级联规模预测更为有效。子图 (a)、子图 (b) 和子图 (d) 中的信息级联表示更为平滑和难以分辨，它们也更有可能遇到负面迁移问题。子图 (c) 中的表示来自于监督学习模型 Base，图中的规模大的信息级联表示并没有分的很开，其较差的效果背后的原因来自于缺少无监督预训练和模型知识蒸馏。

4.9.6.2 模型大小的影响

在表4-12中，我们做了一个额外的有关模型大小消融实验。我们使用带有 1% 标签数据或 100% 标签数据的微博数据集。从表中可知，对于监督学习模型，当模型的大小变得很大的时候，它们的预测效果急剧下降。另一方面，基于预训练和微调的半监督学习模型的预测更为稳定，在模型的大小变大之后，预测效果甚至会提升。

4.9.6.3 在其他数据集上的预测结果

在表4-13中，我们使用知识蒸馏后的 CCGL 模型在推特、ACM、APS 和 DBLP 数据集上的进行信息级联规模预测。我们提出的 CCGL 模型与基于监督学习的 Base 模型相比，在所有的数据集上都与 Base 模型相同或更好。当标签数据的数量受限时，CCGL 模型取得的效果提升更大。

表 4-13 CCGL 模型与基准模型在推特、ACM、APS 和 DBLP 数据集上的效果对比。我们运行五次实验并报告平均均方对数误差（MSLE）及其标准差。我们使用三种不同的标签比例（1%, 10%, 100%）。我们使用自蒸馏的 CCGL 模型，它同时在有标签数据和无标签数据上预训练 30 个轮次。

模型	**推特 (1%)**	**推特 (10%)**	**推特 (Full)**
Feature	-	$7.97_{\pm 0.32}$	$6.78_{\pm 0.52}$
Base	$8.32_{\pm 0.28}$	$6.31_{\pm 0.15}$	$5.32_{\pm 0.05}$
CCGL-FT	$7.49_{\pm 0.23}$	$5.81_{\pm 0.06}$	$5.20_{\pm 0.05}$
CCGL	$\mathbf{7.03}_{\pm 0.06}$	$\mathbf{5.75}_{\pm 0.04}$	$\mathbf{5.12}_{\pm 0.02}$
（效果提升）	15.5%↑	8.9%↑	3.8%↑

模型	**ACM (1%)**	**ACM (10%)**	**ACM (Full)**
Feature	$1.41_{\pm 0.02}$	$1.30_{\pm 0.01}$	$1.25_{\pm 0.01}$
Base	$1.47_{\pm 0.00}$	$1.27_{\pm 0.00}$	$1.23_{\pm 0.01}$
CCGL-FT	$1.42_{\pm 0.01}$	$1.26_{\pm 0.01}$	$\mathbf{1.21}_{\pm 0.00}$
CCGL	$\mathbf{1.42}_{\pm 0.00}$	$\mathbf{1.24}_{\pm 0.00}$	$\mathbf{1.22}_{\pm 0.00}$
（效果提升）	3.4%↑	2.4%↑	0.1%↑

模型	**APS (1%)**	**APS (10%)**	**APS (Full)**
Feature	$2.42_{\pm 0.11}$	$1.88_{\pm 0.01}$	$1.82_{\pm 0.01}$
Base	$2.53_{\pm 0.03}$	$1.85_{\pm 0.03}$	$\mathbf{1.78}_{\pm 0.01}$
CCGL-FT	$2.19_{\pm 0.04}$	$1.88_{\pm 0.02}$	$1.80_{\pm 0.01}$
CCGL	$\mathbf{2.08}_{\pm 0.02}$	$\mathbf{1.81}_{\pm 0.00}$	$\mathbf{1.78}_{\pm 0.00}$
（效果提升）	17.8%↑	2.1%↑	-

模型	**DBLP (1%)**	**DBLP (10%)**	**DBLP (Full)**
Feature	$2.31_{\pm 0.12}$	$2.12_{\pm 0.02}$	$2.01_{\pm 0.03}$
Base	$2.57_{\pm 0.02}$	$2.08_{\pm 0.03}$	$\mathbf{1.93}_{\pm 0.02}$
CCGL-FT	$2.29_{\pm 0.04}$	$2.03_{\pm 0.03}$	$1.94_{\pm 0.02}$
CCGL	$\mathbf{2.18}_{\pm 0.02}$	$\mathbf{2.00}_{\pm 0.02}$	$\mathbf{1.92}_{\pm 0.00}$
（效果提升）	15.2%↑	3.8%↑	0.5%↑

4.10 本章小节

在本章中，我们提出了基于图对比自监督学习的 CCGL 模型，它提供了一种新的对信息级联进行建模的视角和方法。CCGL 模型结合了基于监督学习和无监督学习的信息级联建模与预测，并且支持有效的信息级联图数据增强策略。在五个公开的大规模信息级联数据集上，与强大的监督和半监督学习基准模型相比，

CCGL 取得了非常好的预测效果，这得益于我们所设计的信息级联图数据增强策略和 CCGL 模型的对比自监督预训练、模型微调、知识蒸馏范式。除了效果提升，我们的模型还可以利用大量的无标签数据，从之中抽取有用的知识并利用到自监督预训练之中。我们还展示了 CCGL 模型高效的标签利用性能，和它在不同的信息级联数据集和不同的信息级联预测任务上的泛化性能和迁移性能。

第五章 全文总结与展望

5.1 全文总结

信息传播研究和信息级联建模在近些年来得到了各个研究领域的广泛关注，例如信息决策系统、社交网络分析、图数据学习等[11]。理解信息级联在网络中的传播机制拥有着非常重要的经济影响和社会效益，例如，在最近爆发的新冠病毒疫情中预测一个区域中受到感染的病例及死亡人数对政府和医疗部门作出重要的决策具有关键的意义。

本文以信息传播及信息级联建模为研究背景，主要从信息级联图的时序特征和结构特征作为切入点，使用图神经网络来对信息级联的规模预测问题进行研究、分析、建模和预测。针对信息级联预测领域当前存在的多个重要问题和挑战，我们在本文中提出了两个创新的预测模型：CasFlow 和 CCGL。

针对信息级联预测中缺少层级建模和不确定性建模的问题，我们提出了基于图神经网络的 CasFlow 模型，它从信息级联结构的局部角度和全局角度，使用图小波和稀疏矩阵分解技术来进行图结构的表示学习。CasFlow 模型引入了门控循环单元和变分自编码器来进行层级的信息级联建模和变分推断，从而可以捕获信息在传播中的变化和不确定性。为了克服变分自编码器所假设的简单高斯分布族，我们还引入了正则化流来学习更为灵活和复杂的信息级联后验分布。在与当前最先进的基准模型的对比中，CasFlow 模型不仅取得了最好的预测效果，其学习到的表示还具有很好的可解释性。

针对信息级联预测中存在的无法利用无标签数据、容易陷入过拟合、过于依赖标签数据等问题，我们提出了基于图对比自监督学习的 CCGL 模型，它首先利用了无标签数据来进行不基于特定任务的对比自监督预训练，然后在特定下游任务上进行模型微调和知识蒸馏来提升预测的效果。我们创新地针对信息级联图设计了图数据增强策略 AugSIM，其可以有效地模拟信息在网络中的传播过程。与传统的监督学习模型相比，CCGL 模型显著地提升了信息级联规模预测的效果。我们还在多个信息级联数据集和预测任务上进行了迁移学习实验，实验结果表明 CCGL 模型具有很好的迁移性能和泛化性能。

5.2 后续工作展望

我们计划将 CasFlow 模型扩展到其他类型的特征上，例如，对各种内容特征（文本、图像、论文的标题和摘要特征、以及各类与用户有关的特征，例如用户

的关注数、作者的 h 指数，历史发表文章等）来进行学习。此外，还可以使用更为强大和复杂的图神经网络，例如，拥有多种节点类型和边类型的异构信息网络（Heterogeneous Information Network, HIN），动态图神经网络（图会随着时间变化而变化）等。其他类型的变分推断方法和正则化流方法[137]也可以整合到 CasFlow 模型中去学习更为高阶的隐藏表示以及更为丰富的信息级联后验分布。CasFlow 模型不基于特定的特征、平台和数据，它可以泛化到其他类型的基于图的商业应用领域，例如广告投放策略、市场营销、病毒式信息扩散、可解释的信息预测、谣言检测、流行病控制等。

我们计划在以下几个方面对 CCGL 模型进行扩展和探索：（1）结合多个数据集的无监督预训练来进行更好的知识迁移；（2）其他类型的信息级联图增强策略，例如引入更多的信息级联特征来引导添加节点或边、删除节点或边的过程；（3）使用其他的对比机制，例如动量更新[118]、自回归建模[123]、互信息最大化[122]、多视角对比[131]等；（4）从多模态的角度来进行信息级联表示学习[11]。

致 谢

在攻读硕士学位期间，首先衷心感谢我的两位导师：电子科技大学的钟婷副教授和周帆副教授。在三年多的时间里，他们给我提供了难得的学习机会，对我进行了严谨的科研训练，在各方面对我进行帮助，鼓励我不断努力进取，克服生活和科研上的各种困难。

此外，我还要感谢爱荷华州立大学的 Goce Trajcevski 副教授和马里兰大学帕克分校的 Kunpeng Zhang 助理教授对我在科研上的指导和帮助。感谢实验室的师兄师姐们：Xueqin Chen、Xiaoyan Li、Fang Liu、Xin Liu、Yuhua Mo、Hui Peng、Zijing Wen、Bangying Wu、Xiaoli Yue 等。感谢和我一届的六个优秀的小伙伴：Chengtai Cao、Liang Li、Xiuxiu Qi、Tianliang Wang、Qing Yang、Shengming Zhang。感谢室友 Shijia Cai 和 Yuding Zuo。感谢师弟师妹们：Shupei Chen、Xiaodie Chen、Zhangtao Cheng、Yurou Dai、Wenyue Deng、Xin Jing、Ce Li、Rongfan Li、Wenxiong Li、Lanxing Tuo、Guanyu Wang、Pengyu Wang、Xueting Wang、Zhiyuan Wang、Zheyang Xu、Liu Yu、Weifeng Zhang 等。我所能取得的任何成就都离不开各位老师和同学们的悉心帮助，谢谢你们，祝你们身体健康、前程似锦、诸事顺利。

感谢答辩委员会的专家们对本文的评审和修改建议。

最后感谢我的母亲李岩和女朋友钟文杰，谢谢你们的陪伴与支持。♡♡♡

2015 年，四川成都，电子科技大学，清水河校区（徐增/摄）

参考文献

[1] J. Leskovec, L. Adamic, B. A. Huberman. The dynamics of viral marketing[J]. ACM Transactions on the Web, 2007, 1(1): 5

[2] D. Kempe, J. Kleinberg, É. Tardos. Maximizing the spread of influence through a social network[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2003, 137-146

[3] D. Wang, C. Song, A.-L. Barabási. Quantifying long-term scientific impact[J]. Science, 2013, 342(6154): 127-132

[4] Q. Wu, Y. Gao, X. Gao, et al. Dual sequential prediction models linking sequential recommendation and information dissemination[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2019, 447-457

[5] R. M. Bond, C. J. Fariss, J. J. Jones, et al. A 61-million-person experiment in social influence and political mobilization[J]. Nature, 2012, 489(7415)

[6] L. Zhao, J. Chen, F. Chen, et al. Online flu epidemiological deep modeling on disease contact network[J]. GeoInformatica, 2019, 1-33

[7] R. Kobayashi, R. Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics[C]. International Conference on Web and Social Media (ICWSM), 2016, 191-200

[8] S. Mishra, M.-A. Rizoiu, L. Xie. Feature driven and point process approaches for popularity prediction[C]. International Conference on Information and Knowledge Management (CIKM), 2016, 1069-1078

[9] F. Chen, W. H. Tan, et al. Marked self-exciting point process modelling of information diffusion on Twitter[J]. Annals of Applied Statistics, 2018, 12(4): 2175-2196

[10] Q. Zhao, M. A. Erdogdu, H. Y. He, et al. Seismic: A self-exciting point process model for predicting tweet popularity[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2015, 1513-1522

[11] F. Zhou, X. Xu, G. Trajcevski, et al. A survey of information cascade analysis: Models, predictions, and recent advances[J]. ACM Computing Surveys (CSUR), 2021, 54(2): 1-36

[12] J. Cheng, L. Adamic, P. A. Dow, et al. Can cascades be predicted?[C]. The Web Conference (WWW), 2014, 925-936

[13] X. Cheng, J. Liu, C. Dale. Understanding the characteristics of internet short video sharing: A youtube-based measurement study[J]. IEEE Transactions on Multimedia (TMM), 2013, 15(5): 1184-1194

[14] G. Gürsun, M. Crovella, I. Matta. Describing and forecasting video access patterns[C]. Internation Conference on Computer Communications (INFOCOM), 2011, 16-20

[15] S. Wu, M.-A. Rizoiu, L. Xie. Estimating attention flow in online video network[J]. Proceedings of the ACM on Human-Computer Interaction (CSCW), 2019, 1-25

[16] A. Oghina, M. Breuss, M. Tsagkias, et al. Predicting IMDB movie ratings using social media[C]. European Conference on Information Retrieval (ECIR), 2012, 503-507

[17] K. Lerman, T. Hogg. Using a model of social dynamics to predict popularity of news[C]. The Web Conference (WWW), 2010, 621-630

[18] A. Tatar, P. Antoniadis, M. D. De Amorim, et al. From popularity prediction to ranking online news[J]. Social Network Analysis and Mining (SNAM), 2014, 4(1): 174

[19] J. Qiu, J. Tang, H. Ma, et al. Deepinf: Social influence prediction with deep learning[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2018, 2110-2119

[20] M.-A. Rizoiu, S. Mishra, Q. Kong, et al. Sir-hawkes: Linking epidemic models and hawkes processes to model diffusions in finite populations[C]. The Web Conference (WWW), 2018, 419-428

[21] Q. Cao, H. Shen, K. Cen, et al. Deephawkes: Bridging the gap between prediction and understanding of information cascades[C]. International Conference on Information and Knowledge Management (CIKM), 2017, 1149-1158

[22] Y. Wang, H. Shen, S. Liu, et al. Cascade dynamics modeling with attention-based recurrent neural network[C]. International Joint Conference on Artificial Intelligence (IJCAI), 2017, 2985-2991

[23] A. Tatar, M. D. De Amorim, S. Fdida, et al. A survey on predicting the popularity of web content[J]. Journal of Internet Services and Applications, 2014, 5(1): 8

[24] S. Gao, J. Ma, Z. Chen. Effective and effortless features for popularity prediction in microblogging network[C]. The Web Conference (WWW), 2014, 269-270

[25] J. Yang, J. Leskovec. Patterns of temporal variation in online media[C]. ACM International Conference on Web Search and Data Mining (WSDM), 2011, 177-186

[26] G. Szabo, B. A. Huberman. Predicting the popularity of online content[J]. Communications of the ACM, 2010, 53(8): 80-88

[27] H. Lakkaraju, J. Ajmera. Attention prediction on social media brand pages[C]. International Conference on Information and Knowledge Management (CIKM), 2011, 2157-2160

[28] Y. Matsubara, Y. Sakurai, B. A. Prakash, et al. Rise and fall patterns of information diffusion: model and implications[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2012, 6-14

[29] M. Tsagkias, W. Weerkamp, M. De Rijke. News comments: Exploring, modeling, and online prediction[C]. European Conference on Information Retrieval (ECIR), 2010, 191-203

[30] S. Petrovic, M. Osborne, V. Lavrenko. Rt to win! predicting message propagation in twitter[C]. International Conference on Web and Social Media (ICWSM), 2011

[31] S. Gao, J. Ma, Z. Chen. Popularity prediction in microblogging network[C]. Asia-Pacific Web Conference (APWeb), 2014, 379-390

[32] B. Wu, T. Mei, W.-H. Cheng, et al. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition[C]. AAAI Conference on Artificial Intelligence (AAAI), 2016, 272-278

[33] J. Leskovec, L. Backstrom, J. Kleinberg. Meme-tracking and the dynamics of the news cycle[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2009, 497-506

[34] S. Asur, B. A. Huberman, G. Szabo, et al. Trends in social media: Persistence and decay[C]. International Conference on Web and Social Media (ICWSM), 2011, 1-12

[35] T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning: data mining, inference, and prediction[M]. Springer Science & Business Media, 2009

[36] D. Xie, J. Xu, T.-C. Lu. What's trending tomorrow, today: Using early adopters to discover popular posts on tumblr[C]. IEEE International Conference on Big Data, 2017, 2168-2176

[37] L. Weng, F. Menczer, Y.-Y. Ahn. Predicting successful memes using network and community structure[C]. International Conference on Web and Social Media (ICWSM), 2014

[38] W. Galuba, K. Aberer, D. Chakraborty, et al. Outtweeting the twitterers-predicting information cascades in microblogs[J]. Workshop on Online Social Networks, 2010, 10: 3-11

[39] P. Bao, H.-W. Shen, J. Huang, et al. Popularity prediction in microblogging network: a case study on sina weibo[C]. The Web Conference (WWW), 2013, 177-178

[40] T. R. Zaman, E. B. Fox, E. T. Bradlow, et al. A bayesian approach for predicting the popularity of tweets[J]. The Annals of Applied Statistics, 2014, 8(3): 1583-1611

[41] B. Suh, L. Hong, P. Pirolli, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network[C]. IEEE International Conference on Social Computing (SocialCom), 2010, 177-184

[42] M. Jenders, G. Kasneci, F. Naumann. Analyzing and predicting viral tweets[C]. The Web Conference (WWW), 2013, 657-664

[43] E. Bakshy, J. M. Hofman, W. A. Mason, et al. Everyone's an influencer: quantifying influence on twitter[C]. ACM International Conference on Web Search and Data Mining (WSDM), 2011, 65-74

[44] P. A. Dow, L. A. Adamic, A. Friggeri. The anatomy of large facebook cascades[C]. International Conference on Web and Social Media (ICWSM), 2013, 145-154

[45] N. J. Yuan, Y. Zhong, F. Zhang, et al. Who will reply to/retweet this tweet?: The dynamics of intimacy from online social interactions[C]. ACM International Conference on Web Search and Data Mining (WSDM), 2016, 3-12

[46] L. Hong, O. Dan, B. D. Davison. Predicting popular messages in twitter[C]. Companion of The Web Conference, 2011, 57-58

[47] Z. Yang, J. Guo, K. Cai, et al. Understanding retweeting behaviors in social networks[C]. International Conference on Information and Knowledge Management (CIKM), 2010, 1633-1636

[48] B. Shulman, A. Sharma, D. Cosley. Predictability of popularity: Gaps between prediction and understanding[C]. International Conference on Web and Social Media (ICWSM), 2016, 348-357

[49] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent dirichlet allocation[J]. Journal of Machine Learning Research (JMLR), 2003, 3(Jan): 993-1022

[50] E. Khabiri, C.-F. Hsu, J. Caverlee. Analyzing and predicting community preference of socially generated metadata: A case study on comments in the digg community[C]. International Conference on Web and Social Media (ICWSM), 2009

[51] P. J. McParlane, Y. Moshfeghi, J. M. Jose. Nobody comes here anymore, it's too crowded; predicting image popularity on flickr[C]. International Conference on Multimedia Retrieval (ICMR), 2014, 385

[52] A. Khosla, A. Das Sarma, R. Hamid. What makes an image popular?[C]. The Web Conference (WWW), 2014, 867-876

[53] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks[C]. Annual Conference on Neural Information Processing Systems (NIPS), 2012, 1097-1105

[54] R. Guo, P. Shakarian. A comparison of methods for cascade prediction[C]. International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, 591-598

[55] H. Shen, D. Wang, C. Song, et al. Modeling and predicting popularity dynamics via reinforced poisson processes[C]. AAAI Conference on Artificial Intelligence (AAAI), 2014, 291-297

[56] S. Gao, J. Ma, Z. Chen. Modeling and predicting retweeting dynamics on microblogging platforms[C]. ACM International Conference on Web Search and Data Mining (WSDM), 2015, 107-116

[57] J. G. Lee, S. Moon, K. Salamatian. An approach to model and predict the popularity of online contents with explanatory factors[C]. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010, 623-630

[58] J. G. Lee, S. Moon, K. Salamatian. Modeling and predicting the popularity of online contents with cox proportional hazard regression model[J]. Neurocomputing, 2012, 76(1): 134-145

[59] D. R. Cox. Regression models and life-tables[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1972, 34(2): 187-202

[60] D. Q. Vu, A. U. Asuncion, H. D. R, et al. Dynamic egocentric models for citation networks[C]. International Conference on Machine Learning (ICML), 2011, 857-864

[61] L. Yu, P. Cui, F. Wang, et al. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics[C]. International Conference on Data Mining (ICDM), 2015, 559-568

[62] R. Crane, D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system[J]. Proceedings of the National Academy of Sciences of the USA (PNAS), 2008, 105(41): 15649-15653

[63] W. Ding, Y. Shang, L. Guo, et al. Video popularity prediction by sentiment propagation via implicit network[C]. International Conference on Information and Knowledge Management (CIKM), 2015, 1621-1630

[64] 张志扬, 张凤荔, 陈学勤等. 基于分层注意力的信息级联预测模型 [J]. 计算机科学, 2020, 46(6): 210-209

[65] C. Li, J. Ma, X. Guo, et al. Deepcas: An end-to-end predictor of information cascades[C]. The Web Conference (WWW), 2017, 577-586

[66] B. Perozzi, R. Al-Rfou, S. Skiena. Deepwalk: Online learning of social representations[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2014

[67] J. Chung, C. Gulcehre, K. H. Cho, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv:1412.3555, 2014

[68] D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate[C]. International Conference on Learning Representations (ICLR), 2015, 1-15

[69] C. Li, X. Guo, Q. Mei. Joint modeling of text and networks for cascade prediction[C]. International Conference on Web and Social Media (ICWSM), 2018, 640-643

[70] G. Chen, Q. Kong, W. Mao. An attention-based neural popularity prediction model for social media events[C]. IEEE International Conference on Intelligence and Security Informatics, 2017, 161-163

[71] J. Pennington, R. Socher, C. D. Manning. Glove: Global vectors for word representation[C]. Conference on Empirical Methods in Natural Language Processing, 2014, 1532-1543

[72] A. Grover, J. Leskovec. node2vec: Scalable feature learning for networks[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2016, 855-864

[73] B. Wu, W.-H. Cheng, Y. Zhang, et al. Sequential prediction of social media popularity with deep temporal context networks[C]. International Joint Conference on Artificial Intelligence (IJCAI), 2017, 3062-3068

[74] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770-778

[75] S. Hochreiter, J. Schmidhuber. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780

[76] X. Chen, F. Zhou, K. Zhang, et al. Information diffusion prediction via recurrent cascades convolution[C]. International Conference on Data Engineering (ICDE), 2019, 770-781

[77] Q. Cao, H. Shen, J. Gao, et al. Popularity prediction on social platforms with coupled graph neural networks[C]. ACM International Conference on Web Search and Data Mining (WSDM), 2020, 70-78

[78] F. Zhou, X. Xu, K. Zhang, et al. Variational information diffusion for probabilistic cascades prediction[C]. Internation Conference on Computer Communications (INFOCOM), 2020, 1618-1627

[79] 张志扬, 张凤荔, 谭琪等. 基于深度学习的信息级联预测方法综述 [J]. 计算机科学, 2020, 47(7): 141-153

[80] Z. T. Kefato, N. Sheikh, L. Bahri, et al. Cas2vec: Network-agnostic cascade prediction in online social networks[C]. International Conference on Social Networks Analysis, Management and Security (SNAMS), 2018, 72-79

[81] P. Cui, S. Jin, L. Yu, et al. Cascading outbreak prediction in networks: a data-driven approach[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2013, 901-909

[82] C. Gou, H. Shen, P. Du, et al. Learning sequential features for cascade outbreak prediction[J]. Knowledge and Information Systems, 2018, 57(3): 721-739

[83] D. Liao, J. Xu, G. Li, et al. Popularity prediction on online articles with deep fusion of temporal process and content features[C]. AAAI Conference on Artificial Intelligence (AAAI), 2019, 200-207

[84] M. Tsagkias, W. Weerkamp, M. De Rijke. Predicting the volume of comments on online news stories[C]. International Conference on Information and Knowledge Management (CIKM), 2009, 1765-1768

[85] N. Naveed, T. Gottron, J. Kunegis, et al. Bad news travel fast: A content-based analysis of interestingness on twitter[C]. ACM Conference on Web Science (WebSci), 2011, 1-7

[86] A. Kupavskii, L. Ostroumova, A. Umnov, et al. Prediction of retweet cascade size over time[C]. International Conference on Information and Knowledge Management (CIKM), 2012, 2335-2338

[87] L. Weng, F. Menczer, Y.-Y. Ahn. Virality prediction and community structure in social networks[J]. Scientific Reports, 2013, 3: 2522

[88] H. Pinto, J. M. Almeida, M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos[C]. ACM International Conference on Web Search and Data Mining (WSDM), 2013, 365-374

[89] X. Gao, Z. Cao, S. Li, et al. Taxonomy and evaluation for microblog popularity prediction[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2019, 13(2): 15

[90] W. Zhang, W. Wang, J. Wang, et al. User-guided hierarchical attention network for multi-modal social image popularity prediction[C]. The Web Conference (WWW), 2018, 1277-1286

[91] J. Wang, V. W. Zheng, Z. Liu, et al. Topological recurrent neural network for diffusion prediction[C]. International Conference on Data Mining (ICDM), 2017, 475-484

[92] C. Yang, J. Tang, M. Sun, et al. Multi-scale information diffusion prediction with reinforced recurrent networks[C]. International Joint Conference on Artificial Intelligence (IJCAI), 2019, 4033-4039

[93] X. Chen, K. Zhang, F. Zhou, et al. Information cascades modeling via deep multi-task learning[C]. International ACM SIGIR conference on research and development in Information Retrieval (SIGIR), 2019, 885-888

[94] D. K. Hammond, P. Vandergheynst, R. Gribonval. Wavelets on graphs via spectral graph theory[J]. Applied and Computational Harmonic Analysis, 2011, 30(2): 129-150

[95] D. I. Shuman, S. K. Narang, P. Frossard, et al. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains[J]. IEEE Signal Processing Magazine, 2013, 30(3): 83-98

[96] C. Donnat, M. Zitnik, D. Hallac, et al. Learning structural node embeddings via diffusion wavelets[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), London, UK, 2018, 1320-1329

[97] E. Lukacs. Characteristic functions[M]. London, UK: Griffin, 1970

[98] J. Zhang, Y. Dong, Y. Wang, et al. ProNE: fast and scalable network representation learning[C]. International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 2019, 4278-4284

[99] T. Tao. Topics in random matrix theory[M]. American Mathematical Society, 2012

[100] N. Halko, P.-G. Martinsson, J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions[J]. SIAM Review, 2011, 53(2): 217-288

[101] S. Lamprier. A recurrent neural cascade-based model for continuous-time diffusion[C]. International Conference on Machine Learning (ICML), Long Beach, California, USA, 2019, 1-10

[102] D. P. Kingma, M. Welling. Auto-encoding variational bayes[C]. International Conference on Learning Representations (ICLR), 2014

[103] W. Xu, H. Sun, C. Deng, et al. Variational autoencoder for semi-supervised text classification[C]. AAAI Conference on Artificial Intelligence (AAAI), Francisco, California, USA, 2017, 3358-3364

[104] D. Zhu, P. Cui, D. Wang, et al. Deep variational network embedding in Wasserstein space[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), London, UK, 2018, 2827 -2836

[105] Q. Gao, F. Zhou, G. Trajcevski, et al. Predicting human mobility via variational attention[C]. The Web Conference (WWW), San Francisco, California, USA, 2019, 2750-2756

[106] D. Liang, R. G. Krishnan, M. D. Hoffman, et al. Variational autoencoders for collaborative filtering[C]. The Web Conference (WWW), Lyon, France, 2018, 689-698

[107] M. Cha, H. Kwak, P. Rodriguez, et al. Analyzing the video popularity characteristics of large-scale user generated content systems[J]. IEEE/ACM Transactions on Networking (TON), 2009, 17(5): 1357-1370

[108] D. Rezende, S. Mohamed. Variational inference with normalizing flows[C]. International Conference on Machine Learning (ICML), Lille, France, 2015, 1530-1538

[109] L. Dinh, J. Sohl-Dickstein, S. Bengio. Density estimation using Real NVP[C]. International Conference on Learning Representations (ICLR), Toulon, France, 2017

[110] D. P. Kingma, T. Salimans, R. Jozefowicz, et al. Improved variational inference with inverse autoregressive flow[C]. Annual Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, 2016, 4743-4751

[111] D. I. Shuman, P. Vandergheynst, P. Frossard. Chebyshev polynomial approximation for distributed signal processing[C]. International Conference on Distributed Computing in Sensor Systems (DCOSS), Barcelona, Spain, 2011, 1-8

[112] M. Defferrard, X. Bresson, P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering[C]. Annual Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, 2016, 3837-3845

[113] S. Goel, A. Anderson, J. Hofman, et al. The structural virality of online diffusion[J]. Management Science, 2015, 62(1): 180-196

[114] L. Van Der Maaten. Accelerating t-SNE using tree-based algorithms[J]. Journal of Machine Learning Research (JMLR), 2014, 15(1): 3221-3245

[115] M. Wu, S. Pan, C. Zhou, et al. Unsupervised domain adaptive graph convolutional networks[C]. The Web Conference (WWW), 2020, 1457-1467

[116] W. Hu, B. Liu, J. Gomes, et al. Strategies for pre-training graph neural networks[C]. International Conference on Learning Representations (ICLR), 2020, 1-22

[117] X. Gao, X. Jia, C. Yang, et al. Using survival theory in early pattern detection for viral cascades[J]. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2020

[118] K. He, H. Fan, Y. Wu, et al. Momentum contrast for unsupervised visual representation learning[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 9729-9738

[119] Z. Wu, Y. Xiong, S. X. Yu, et al. Unsupervised feature learning via non-parametric instance discrimination[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 3733-3742

[120] I. Misra, L. v. d. Maaten. Self-supervised learning of pretext-invariant representations[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 6707-6717

[121] P. Bachman, R. D. Hjelm, W. Buchwalter. Learning representations by maximizing mutual information across views[C]. Annual Conference on Neural Information Processing Systems (NeurIPS), 2019, 15535-15545

[122] P. Velickovic, W. Fedus, W. L. Hamilton, et al. Deep graph infomax[C]. International Conference on Learning Representations (ICLR), 2019, 1-17

[123] A. v. d. Oord, Y. Li, O. Vinyals. Representation learning with contrastive predictive coding[J]. arXiv:1807.03748, 2018

[124] T. Chen, S. Kornblith, M. Norouzi, et al. A simple framework for contrastive learning of visual representations[J]. International Conference on Machine Learning (ICML), 2020, 1-18

[125] T. Chen, S. Kornblith, K. Swersky, et al. Big self-supervised models are strong semi-supervised learners[C]. Annual Conference on Neural Information Processing Systems (NeurIPS), 2020

[126] E. D. Cubuk, B. Zoph, D. Mane, et al. Autoaugment: Learning augmentation strategies from data[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 113-123

[127] T. Zhao, Y. Liu, L. Neves, et al. Data augmentation for graph neural networks[J]. arXiv:2006.06830, 2020

[128] Y. Rong, W. Huang, T. Xu, et al. Dropedge: Towards deep graph convolutional networks on node classification[C]. International Conference on Learning Representations (ICLR), 2019, 1-17

[129] D. Chen, Y. Lin, W. Li, et al. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view[C]. AAAI Conference on Artificial Intelligence (AAAI), 2020, 3438-3445

[130] J. Qiu, Q. Chen, Y. Dong, et al. GCC: Graph contrastive coding for graph neural network pre-training[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2020, 1150-1160

[131] K. Hassani, A. H. Khasahmadi. Contrastive multi-view representation learning on graphs[C]. International Conference on Machine Learning (ICML), 2020, 1-13

[132] W. Ying, Y. Zhang, J. Huang, et al. Transfer learning via learning to transfer[C]. International Conference on Machine Learning (ICML), 2018, 5085-5094

[133] Z. Hu, Y. Dong, K. Wang, et al. GPT-GNN: Generative pre-training of graph neural networks[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2020

[134] F.-Y. Sun, J. Hoffman, V. Verma, et al. InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization[C]. International Conference on Learning Representations (ICLR), 2020, 1-15

[135] X. Chen, H. Fan, R. Girshick, et al. Improved baselines with momentum contrastive learning[J]. arXiv:2003.04297, 2020

[136] J. Tang, J. Zhang, L. Yao, et al. ArnetMiner: Extraction and mining of academic social networks[C]. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2008, 990-998

[137] E. Hajiramezanali, A. Hasanzadeh, K. Narayanan, et al. Variational graph recurrent neural networks[C]. Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 2019, 10700-10710

攻读硕士学位期间取得的成果

学术论文

[1] F. Zhou, **X. Xu**, K. Zhang, et al. Variational information diffusion for probabilistic cascades prediction[C]. IEEE International Conference on Computer Communications (INFOCOM), Virtual Conference, Jul 6-9, 2020, 1618-1627

[2] F. Zhou, **X. Xu**, G. Trajcevski, et al. A survey of information cascade analysis: Models, predictions, and recent advances[J]. ACM Computing Surveys (CSUR), 2021, 54(2): 27:1-27:36

[3] F. Zhou, L. Yu, **X. Xu**, et al. Decoupling representation and regressor for long-tailed information cascade prediction[C]. International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Online, Jul 11-15, 2021

[4] F. Zhou, P. Wang, **X. Xu**, et al. Contrastive trajectory learning for tour recommendation[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021

[5] F. Zhou, X. Jing, **X. Xu**, et al. Continual information cascade learning[C]. IEEE Global Communications Conference (GLOBECOM), Virtual Conference, Dec 7-11, 2020, 1-6

[6] F. Zhou, Z. Wen, T. Zhong, et al. Unsupervised user identity linkage via graph neural networks[C]. IEEE Global Communications Conference (GLOBECOM), Virtual Conference, Dec 7-11, 2020, 1-6

[7] F. Zhou, X. Qi, **X. Xu**, et al. Meta-learned user preference for topic participation prediction[C]. IEEE Global Communications Conference (GLOBECOM), Virtual Conference, Dec 7-11, 2020, 1-6

获奖

[1] 徐增. 计算机通信国际会议（INFOCOM）2020 学生会议奖, 2020 年