Computer Science and Application 计算机科学与应用, 2026, 16(6), 117-128
Published Online June 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.166213

Hans汉斯

基于信任门控异构图学习的社交媒体谣言检测方法

王加瑞，董晓芳，杨　凯*

西京学院计算机学院，陕西　西安

收稿日期：2026年5月10日；录用日期：2026年6月15日；发布日期：2026年6月24日

摘　要

社交媒体谣言检测在很大程度上依赖于结构特征。然而，现有图方法在评估不同节点与交互关系的可靠性方面仍存在不足。为解决这一问题，本文提出了一种自适应信任评估框架(Adaptive Trust Evaluation Framework, ATEF)。ATEF将新闻事件建模为帖子级异构图，能够同时刻画扩散关系、反馈关系以及伪时间关系，从而充分保留传播结构信息与时间顺序信息。进一步地，本文引入了一种信任门控异构消息传递机制(Trust-Gated Heterogeneous Message Passing)，用于自适应调节不同节点及不同关系类型在信息传播过程中的贡献。通过该机制，ATEF能够增强关键传播信号，同时抑制噪声信息的干扰。实验结果表明，ATEF具有优异的检测性能。在测试集上，该模型的Accuracy和Macro-F1均达到0.9512，Fake Recall达到0.9401，显著优于BiGCN基线模型。最后，协同攻击实验与消融实验进一步验证了ATEF在复杂且高度扰动环境下具有较强的鲁棒性。

关键词

社交媒体谣言检测，自适应信任评估框架，帖子级异构图，信任门控异构消息传递

Social Media Rumor Detection Method Based on Trust-Gated Heterogeneous Graph Learning

Jiarui Wang, Xiaofang Dong, Kai Yang*

School of Computer, Xijing University, Xi'an Shaanxi

Received: May 10, 2026; accepted: June 15, 2026; published: June 24, 2026

*通讯作者。

文章引用: 王加瑞, 董晓芳, 杨凯. 基于信任门控异构图学习的社交媒体谣言检测方法[J]. 计算机科学与应用, 2026, 16(6): 117-128. DOI: 10.12677/csa.2026.166213

Abstract

Social media rumor detection largely relies on structural features. However, existing graph methods still have limitations in evaluating the reliability of different nodes and interaction relationships. To address this issue, this paper proposes an Adaptive Trust Evaluation Framework (ATEF). ATEF models news events as post-level heterogeneous graphs, which can simultaneously characterize diffusion relationships, feedback relationships, and pseudo-temporal relationships, thereby fully preserving propagation structure information and temporal order information. Furthermore, this paper introduces a Trust-Gated Heterogeneous Message Passing mechanism to adaptively adjust the contributions of different nodes and relationship types in the information propagation process. Through this mechanism, ATEF can enhance key propagation signals while suppressing the interference of noise information. Experimental results show that ATEF has excellent detection performance. On the test set, the model achieves an Accuracy and Macro-F1 of 0.9512, and a Fake Recall of 0.9401, which significantly outperforms the BiGCN baseline model. Finally, collaborative attack experiments and ablation experiments further verify that ATEF has strong robustness in complex and highly perturbed environments.

Keywords

Social Media Rumor Detection, Adaptive Trust Evaluation Framework, Post-Level Heterogeneous Graph, Trust-Gated Heterogeneous Message Passing

1. 引言

社交媒体现已成为人们获取和分享新闻的主要渠道。然而，它也加速了谣言与虚假信息的传播。Vosoughi 等人的大规模研究表明[1]，虚假新闻在社交平台上的传播速度更快、范围更广、影响更深。因此，谣言检测并不仅仅是一个文本分类任务，而本质上是一个与传播结构密切相关的图学习问题。进一步地，已有综述指出，社交媒体谣言检测与传统新闻核查之间存在显著差异。检测结果不仅取决于文本语义，还与用户交互、扩散路径以及上下文关系等社会传播因素密切相关。因此，仅依赖内容特征往往难以获得稳定且鲁棒的检测性能。

早期方法主要依赖文本语义或时间序列上下文进行谣言检测。例如，基于循环神经网络的事件级建模表明[2]，传播过程中的文本动态变化能够提供有价值的判别线索。随后，研究者开始探索结构化传播信息，并将新闻事件表示为传播树或传播图，以通过结构模式识别谣言。Ma 等人证明[3]，传播结构本身在谣言检测中发挥着重要作用。此外，Monti 等人和 Bian 等人分别采用几何深度学习和双向图卷积网络对社交媒体传播行为进行建模[4]-[6]。与传统基于文本的方法相比，这些结构化模型在虚假新闻和谣言检测任务中取得了更好的性能。然而，现有基于图的方法通常依赖相对统一的邻域聚合策略，难以显式区分不同传播节点和交互关系对检测结果的不同贡献。即使引入图注意力机制[7]，其权重本质上仍主要关注局部相关性，尚不足以针对谣言传播场景显式建模“信任”或“可靠性”。此外，已有研究表明[8]，在含噪图结构中对邻居和边进行自适应重加权，能够显著增强模型鲁棒性。这说明，在复杂传播图任务中引入显式的可靠性调节机制具有重要的实际意义。

基于上述观察，本文提出了一种面向社交媒体谣言检测的自适应信任评估框架(Adaptive Trust Evaluation Framework, ATEF)。该设计使模型能够增强关键传播信号，同时抑制噪声干扰。因此，ATEF 能够显著提升谣言检测的整体性能以及虚假新闻识别能力。

本文的主要贡献总结如下：

(1) 本文提出了一种帖子级异构传播图建模方法。该方法将新闻事件表示为包含扩散关系、反馈关系和伪时间关系的异构图，从而较为全面地保留社交媒体传播过程中蕴含的结构信息。

(2) 本文设计了一种信任门控异构消息传递机制。通过在图表示学习过程中显式引入可学习的信任门控，该机制能够自适应调节不同传播节点和不同关系类型的贡献。

2. 关键技术介绍

当前关于社交媒体谣言检测的研究通常可分为三类：基于内容的语义方法、基于结构的图学习方法，以及融合节点可靠性或信任建模的增强型图方法。这些方法分别从文本表达、传播拓扑和节点交互等不同角度处理谣言检测问题，共同推动了谣言检测研究从简单的静态特征分类逐步转向复杂的结构表示学习。已有研究表明[1]，虚假新闻相较于真实信息传播得更快、更深且更广。这说明，仅依赖文本内容难以捕捉谣言传播的真实本质。例如，Ma 等人使用 RNN 表明[2]传播过程中的文本动态变化能够提供有价值的线索。然而，这类方法仍然高度依赖内容信息。因此，当文本稀疏、伪装性较强，或早期传播数据不足时，其检测性能往往会下降。

为了解决基于内容方法的局限性，研究者逐渐将关注点转向结构化传播信息。他们利用转发、评论和回复等交互行为，将谣言检测建模为结构学习任务。Ma 等人进一步提出了传播树核方法[3]，证明传播结构本身能够为谣言识别提供有效线索。在此基础上，图神经网络(Graph Neural Networks, GNNs)被引入该领域。Monti 等人应用几何深度学习[4]，通过融合传播信息、用户活动和社交网络数据来检测虚假新闻。此外，Bian 等人提出了 BiGCN 模型[5]。该模型通过同时捕获自顶向下的传播路径和自底向上的反馈模式，实现了显著的性能提升。

近年来，研究者开始探索在图学习中引入邻居加权、边重加权和异常抑制等机制。这些机制旨在提升模型在复杂含噪环境中的适应能力。例如，图注意力网络(Graph Attention Networks, GAT)通过学习邻居权重改进了标准图卷积[7]，异构图注意力网络则进一步通过节点级和语义级注意力刻画不同节点及关系语义的重要性差异[8]。然而，这些权重主要捕获局部特征相关性，并未在谣言传播语境下显式建模“信任”或“可靠性”。类似地，GNNGuard 等框架表明[9]，对边进行自适应重加权能够显著增强模型抵抗图噪声与扰动的鲁棒性。因此，一个关键挑战仍然存在：如何将自适应信任调节机制显式嵌入图表示学习过程。这样可以使模型更细粒度地区分有价值的传播证据与干扰性噪声。

3. 系统分析

3.1. 异构事件图表示

对于每一条新闻事件，本文构建一个事件级传播图

$$G_e = \left(V_e, E_e, R_e, X_e\right) \tag{1}$$

其中 V_e 表示事件 e 中的帖子节点集合，E_e 表示节点之间的传播边集合，R_e 表示边类型集合，X_e 表示节点特征矩阵。与以用户为中心的二部图建模不同，本文采用帖子级建模策略，即每个节点对应一次具体的内容发布、转发、回复或引用行为，而非跨事件复用的用户身份。这样做的原因在于，社交媒体传播数据通常更容易稳定地提供帖子层面的关系信息，而用户跨事件身份及长期行为属性往往难以完整、可

靠地构建。采用帖子级表示后，模型能够更加直接地刻画单条新闻在传播过程中的结构变化与交互模式。

在图结构上，本文引入三类关系来描述传播过程中的异构交互信息，即自上而下的扩散边、自下而上的反馈边以及用于保留局部顺序信息的伪时间相邻边。由此，事件图不再是单一关系下的普通传播树，而是包含多种传播关系的异构传播图，这为后续的信任门控消息传递提供了更丰富的建模基础。

为了充分表征传播节点的局部语义与结构状态，本文为每个帖子节点构造联合特征表示：

$$x_i = \left[x_i^{\text{text}} \,\|\, x_i^{\text{struct}} \,\|\, x_i^{\text{temp}} \right] \tag{2}$$

其中，x_i^{text} 表示文本语义特征，x_i^{struct} 表示结构属性特征，x_i^{temp} 表示时间或顺序相关特征。具体而言，文本特征用于刻画标题或帖子文本中的表达模式与潜在语义倾向，结构特征用于描述节点在传播图中的位置与角色，时间特征则用于反映节点在事件传播过程中的相对顺序。考虑到不同来源特征的维度与分布存在差异，本文首先通过共享的非线性映射将其投影到统一的隐空间中，得到节点初始表示：

$$h_i^{(0)} = \phi\left(W_0 x_i + b_0 \right) \tag{3}$$

其中 W_0 和 b_0 为可学习参数，$\phi(\cdot)$ 表示非线性激活函数。

3.2. 基于信任门控的异构消息传递

ATEF 的核心在于信任门控异构消息传递机制。与传统图神经网络对所有邻居进行统一聚合不同，本文认为不同传播节点对目标节点的贡献并不一致，因此为每个节点引入可学习的信任门值，以刻画其在当前传播上下文中的相对可靠性：

$$g_j^{(l)} = \sigma\left(W_g^{(l)} h_j^{(l)} + b_g^{(l)} \right) \tag{4}$$

其中 $g_j^{(l)} \in (0,1)$，$\sigma(\cdot)$ 为 sigmoid 函数，$W_g^{(l)}$ 和 $b_g^{(l)}$ 为第 l 层的可学习参数。

在第 l 层表示学习中，节点 v_i 的更新过程写为

$$h_i^{(l+1)} = \phi\left(W_s^{(l)} h_i^{(l)} + \sum_{r \in R_e} \sum_{j \in N_r(i)} \alpha_{ij}^{(l,r)} g_j^{(l)} W_r^{(l)} h_j^{(l)} \right) \tag{5}$$

其中，$W_s^{(l)}$ 为自连接映射矩阵，$W_r^{(l)}$ 为关系类型，r 对应的线性映射矩阵，$\mathcal{N}_r(i)$ 表示节点 v_i 在关系 r 下的邻居集合，$\alpha_{ij}^{(l,r)}$ 表示关系感知的注意力权重。该式表明，ATEF 在关系特定映射与局部注意力的基础上，进一步利用 trust gate 对邻居贡献进行自适应调制，从而突出关键传播信号并抑制噪声干扰。

3.3. 图级预测与优化

经过多层信任门控异构消息传递后，本文将所有节点的高层表示聚合为图级表示，并输出事件级真假预测。图级读出与分类过程统一写为

$$\hat{y}_e = \text{softmax}\left(W_c \, Readout\left(H_e^{(L)} \right) + b_c \right) \tag{6}$$

其中，$H_e^{(L)}$ 表示事件图中所有节点在第 L 层的表示集合，$Readout(\cdot)$ 表示图级聚合操作，本文实现中采用均值池化与最大池化相结合的方式，W_c 和 b_c 为分类层参数。

在训练阶段，本文采用分类损失与门控正则项的联合优化目标，$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{gate}$，其中，$\mathcal{L}_{cls}$ 表示标准二分类交叉熵损失，$\mathcal{L}_{gate}$ 表示用于缓解 trust gate 过度饱和的轻量正则项，λ 为平衡系数。通过该联合优化策略，模型既能够学习有利于谣言判别的传播图表示，又能够保持信任门控机制的稳定性。

综合来看，ATEF 的整体流程见图 1。具体可以分为四步流程：首先，针对每条新闻事件构建帖子级异构传播图，并提取文本、结构和时间三类特征；其次，通过共享编码层获得节点初始表示；随后，在多

层 trust-gated heterogeneous message passing 中结合关系特定映射、局部注意力与信任门控完成跨关系消息聚合；最后，通过图级读出与分类层输出事件级真假预测，并利用联合损失函数进行端到端训练。通过将异构传播结构建模、多源特征融合和可学习信任调制统一到同一框架中，ATEF 能够更细粒度地建模不同节点及不同关系的贡献差异，从而在复杂社交媒体环境下获得更稳健的谣言检测效果。

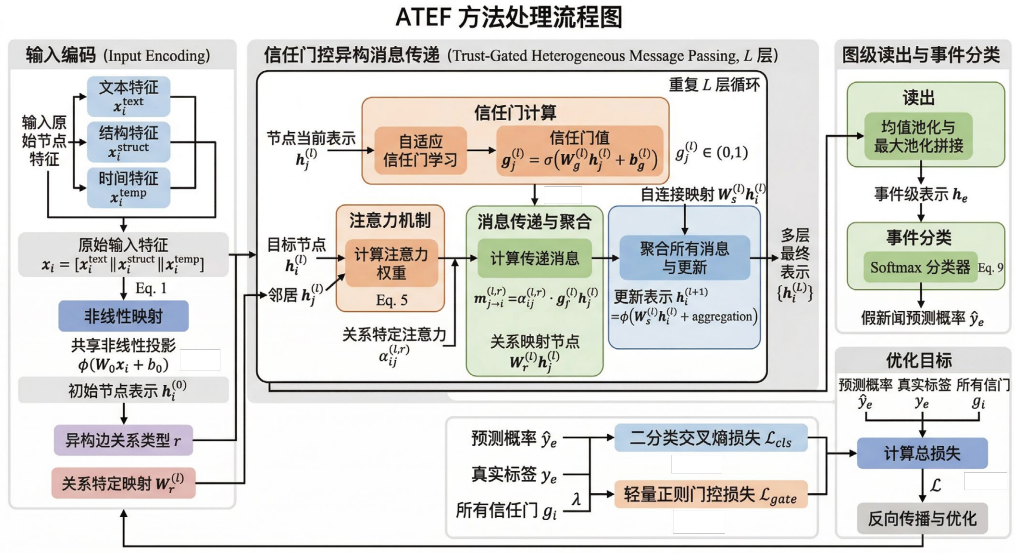

Figure 1. Overall Architecture of ATEF
图 1. ATEF 整体流程图

4. 实验设置

4.1. 参数设置

本文实验采用 FakeNewsNet 中的 GossipCop 子数据集。该数据集由 Shu 等人构建，原始数据来源于 FakeNewsNet 公开仓库[10]，数据集以新闻事件为基本样本，每个事件对应一个传播图，包含源新闻节点及其相关社交传播节点，并提供事件级真实性标签，即 fake 或 real，用于评估模型对虚假新闻的识别能力。

本文在事件级帖子传播图上评估所提出的 ATEF 模型。具体而言，每条新闻事件被建模为一个帖子级异构传播图，图中同时包含自上而下扩散边、自下而上反馈边以及伪时间相邻边，节点特征由文本特征、结构特征和时间特征拼接构成。模型采用两层 trust-gated heterogeneous message passing，隐藏维度设为 128，dropout 设为 0.2，训练轮数为 40，学习率为 0.001，权重衰减系数为 0.0001，随机种子设为 42，并在 cuda 设备上完成训练。数据划分采用训练集、验证集和测试集三部分，对应样本数分别为 1019、505 和 3542。

本文采用 Accuracy、Precision、Recall、F1-score、Macro-F1 以及 AUC 作为主要评价指标全面评估 ATEF 在谣言检测任务中的性能。由于本文关注的是新闻事件级二分类任务，混淆矩阵中的四个基本统计量分别记为：真阳性(TP)、假阳性(FP)、真阴性(TN)和假阴性(FN)。其中，本文默认将假新闻类别视为正类(positive class)，真实新闻类别视为负类(negative class)。

整体分类准确率(Accuracy)用于衡量模型在全部测试样本上的总体判别正确率，其定义为

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

精确率(Precision)用于衡量被模型判定为某一类别的样本中，真正属于该类别的比例。以假新闻类别为例，其精确率定义为

$$\text{Precision}_{\text{fake}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

召回率(Recall)用于衡量某一类别的真实样本中，被模型成功识别出来的比例。对于假新闻类别，其召回率定义为

$$\text{Recall}_{\text{fake}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

F1-score 是 Precision 与 Recall 的调和平均，用于综合衡量模型在某一类别上的分类质量。对于假新闻类别，其定义为

$$\text{Fl}_{\text{fake}} = \text{Precision}_{\text{fake}} + \text{Recall}_{\text{fake}} \tag{10}$$

相比单独使用 Precision 或 Recall，F1-score 更能反映模型在类别识别中的平衡能力。类似地，真实新闻类别的 Precision、Recall 和 F1-score 可按相同方式计算。

为了同时衡量模型在真假新闻两类样本上的整体表现，本文进一步采用 Macro-F1 作为主要综合指标，其定义为

$$\text{Macro-Fl} = \frac{\text{Fl}_{\text{fake}} + \text{Fl}_{\text{real}}}{2} \tag{11}$$

Macro-F1 对每个类别赋予相同权重，因此相比 Accuracy 更适合用于评价类别间性能是否均衡。

此外，本文还采用 ROC 曲线下面积(AUC)评估模型对假新闻类别的排序能力。AUC 不依赖固定阈值，而是衡量模型在不同判定阈值下区分正负样本的整体能力。AUC 越高，说明模型越能够将假新闻样本排在真实新闻样本之前。由于本文主要关注假新闻检测性能，因此实验中重点报告 AUC (fake)。

4.2. 评价指标

本文将原始传播数据构建为用户–新闻异构二部图，并分别从新闻节点、用户节点、边结构以及时间顺序四个方面提取特征。对于新闻节点，本文将根据新闻内容向量与结构统计量进行拼接，形成新闻节点的初始特征表示。其中，结构统计量主要包括参与传播的用户数量和原始传播图中的节点数量。为降低传播规模差异对模型训练的影响，相关结构统计量首先经过 log(1+x)变换，然后进行 z-score 标准化处理。

对于用户节点，本文根据用户历史交互状态构造节点特征，主要包括最终信任值、正向交互计数、负向交互计数、总交互次数以及度先验等信息。其中，最终信任值直接反映用户在传播过程中的相对可靠性，因此保留其原始数值形式，并作为后续信任门控机制的重要输入；其余连续型特征则进行标准化处理，以缓解不同特征量纲差异对模型学习过程的影响。

对于用户–新闻边，本文进一步构造边特征以刻画用户与新闻之间的传播强度。具体而言，边特征包括交互次数、对数交互次数及其归一化形式。考虑到当前数据集中缺少稳定、可靠的真实时间戳，本文基于事件处理顺序构造相对时间特征，包括用户首次参与时间、相对于新闻首次出现时刻的传播延迟，以及归一化延迟比例。上述时间特征被作为增强边特征用于后续消融实验。

通过上述特征构建方式，本文在保留用户–新闻交互结构信息的同时，进一步引入用户可靠性、传播强度以及相对时间顺序等信息，从而使模型能够在有限时间信息条件下刻画事件传播的先后关系，并为信任门控异构消息传递提供更加充分的输入表示。

5. 实验结果分析

5.1. 基线模型对比

为了验证 ATEF 的有效性，本文将其与经典传播结构谣言检测模型双向图卷积网络(BiGCN)进行了对比实验，对比结果见表 1。

Table 1. Comparative performance of ATEF and BiGCN
表 1. ATEF 与 BiGCN 的性能对比

Metric	ATEF	BiGCN
Accuracy	0.9512	0.8602
Macro-F1	0.9512	0.8583
Fake Precision	0.9622	0.9806
Fake Recall	0.9401	0.7373
Fake F1	0.9510	0.8417
Real Precision	0.9405	0.7868
Real Recall	0.9624	0.9852
Real F1	0.9513	0.8749
AUC (fake)	0.9836	0.9825

由表 1 可以看出，ATEF 在大多数指标上均优于 BiGCN，尤其在整体分类性能和假新闻召回能力方面优势更为明显。具体而言，ATEF 的 Accuracy 和 Macro-F1 分别达到 0.9512 和 0.9512，较 BiGCN 分别提升 9.10 和 9.29 个百分点；与此同时，Fake Recall 由 0.7373 显著提升至 0.9401，Fake F1 由 0.8417 提升至 0.9510，说明 ATEF 能够更有效地识别假新闻样本并降低漏检率。

需要指出的是，虽然 BiGCN 在 Fake Precision 上略高于 ATEF，但其明显较低的 Fake Recall 表明该模型在假新闻检测上呈现出更强的保守性，即更倾向于在保证高精度的前提下牺牲部分召回能力。相比之下，ATEF 在保持较高 Precision 的同时显著提升了 Recall，因此在真假新闻两类样本上表现出更均衡、更稳健的分类能力。综合上述结果可知，ATEF 在整体性能与类别平衡性方面均优于对比基线，验证了所提方法的有效性。

5.2. 不同传播规模下的性能表现

为进一步分析 ATEF 性能优势的来源，本文按照传播图规模对测试集样本进行分组，并比较不同传播规模下的假新闻召回率(Fake Recall)。具体而言，本文以事件传播图中的边数作为划分依据，将测试集划分为 1~5、6~10、11~20 和>20 四个传播规模区间，并统计每个区间内 ATEF 与 BiGCN 的假新闻召回率。图中灰色柱状表示各区间对应的测试样本数量。

ATEF 在不同传播规模下均保持了较高且相对稳定的假新闻召回率见图 2。尤其是在传播规模逐渐增大时，ATEF 仍然能够维持较高的召回水平，而 BiGCN 的召回率则随着传播图复杂度的提升出现明显下降。在最小规模区间内，两种方法的性能差距相对有限；但当传播规模进入中等及以上区间后，ATEF 的优势逐步扩大，并在大规模传播图上表现出更好的稳定性。

上述结果表明，ATEF 的性能提升并非仅来源于少量简单样本，而是在不同传播规模条件下均能够保持较为稳定的检测效果。特别是在传播结构更复杂、关系更丰富的样本中，ATEF 仍能有效保留关键传播

模式并抑制噪声信息干扰，从而获得更高的假新闻识别率。相比之下，BiGCN 对传播结构变化更为敏感，其性能在传播图规模增大时更容易出现退化。该实验进一步说明，本文提出的 trust-gated heterogeneous message passing 机制在复杂传播场景下具有更好的适应能力。

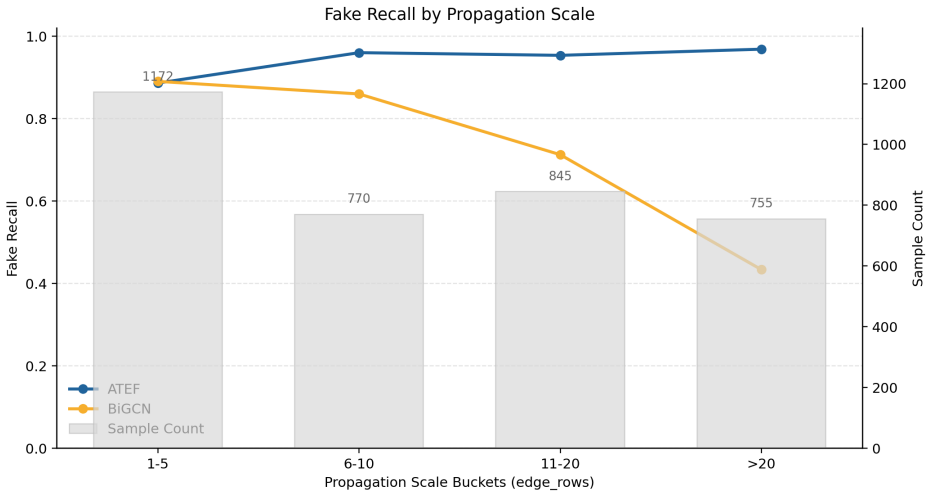

Figure 2. Comparison of fake recall between ATEF and BiGCN under different propagation graph sizes
图 2. 不同传播图规模下 ATEF 与 BiGCN 的 Fake Recall 对比

5.3. 信任门控机制定性与定量分析

已有可解释虚假新闻检测研究通常通过注意力权重、关键评论或可疑转发用户来解释模型判别依据[11][12]；同时，图神经网络解释性研究也常通过识别关键子图结构和重要节点特征来分析模型决策过程[13][14]。受此类工作的启发，本文在测试集上导出最后一层信任门控输出，并从整体统计和个案可视化两个角度分析 trust gate 机制的作用。

本文按照节点在用户－新闻异构图中的结构角色和传播行为，将节点划分为新闻节点、早期参与用户、高交互用户、普通交互用户以及晚期/低活跃用户。其中，新闻节点表示待检测新闻事件；早期参与用户表示归一化传播延迟不超过 0.2 的用户；高交互用户表示在单个事件内部交互次数位于前 20%且交互次数大于 1 的用户；晚期/低活跃用户表示归一化传播延迟不低于 0.8 或交互次数不超过 1 的用户；其余用户划分为普通交互用户，并统计各类节点的平均 trust gate 分数，结果如表 2 所示。不同类型节点的 trust gate 分数整体分布较为集中，均值主要位于 0.4976~0.4980 之间，说明模型并未对节点贡献进行极端化筛选，而是采用较平滑的方式调节不同节点在消息传递中的影响。其中，普通交互用户的平均信任分数相对最高，为 0.497972，但该类节点数量较少，仅占 0.17%，因此其统计结果更适合作为局部现象参考。晚期/低活跃用户和高交互用户的平均信任分数略高于新闻节点与早期参与用户，表明模型在当前数据条件下综合节点交互强度、传播延迟和上下文结构来判断其信息贡献。

本文选取一个典型谣言样本，对其用户－新闻异构传播图中的 trust gate 分数进行可视化展示见图 3。图中方形节点表示新闻源节点，圆形节点表示参与传播的用户节点，节点颜色表示样本内归一化后的 trust gate 分数，颜色越深表示该节点在当前样本中获得的相对门控分数越高。为避免归一化颜色造成误解，图中同时给出了该样本中原始 gate_score 的取值范围。从图中可以看出，不同用户节点获得的 trust gate 分数存在一定差异。部分节点被赋予较高的相对门控分数，说明其在当前传播图上下文中对图表示学习

具有较高的信息贡献；而部分节点的门控分数相对较低，说明模型在消息传递过程中降低了其影响。该结果表明，信任门控机制能够根据节点在传播过程中的结构位置和交互状态对其信息贡献进行自适应调节。

Table 2. Trust gate score statistics across different node types
表 2. 不同类型节点的信任门控分数统计

节点类型	平均信任分数	标准差	节点数量	节点占比
新闻节点	0.497632	0.000410	3542	6.57%
高交互用户	0.497638	0.000752	2463	4.57%
早期参与用户	0.497631	0.000541	23512	43.61%
普通交互用户	0.497972	0.001195	94	0.17%
晚期/低活跃用户	0.497725	0.000492	24307	45.08%

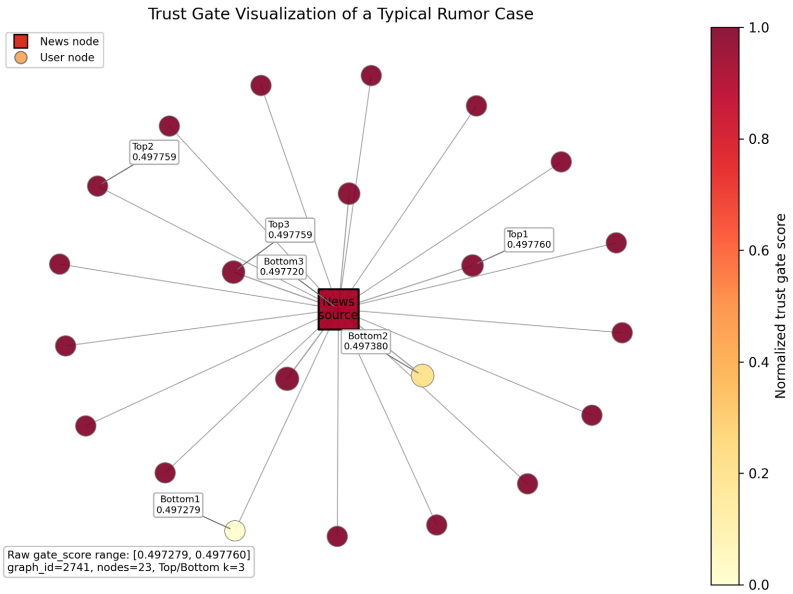

Figure 3. Visualization of trust gate scores in representative rumor samples
图 3. 典型谣言样本中 trust gate 分数的可视化结果

其中，原始 gate_score 为 0.497760 的用户节点，是该事件传播图中的 Top1 高信任节点。节点属性见表 3。该节点的 interaction_count = 2，高于样本中位数 1，并达到样本内 80%分位数水平，说明该节点与新闻节点之间具有相对更强的交互关系。在可视化图中，边宽由交互次数表示，因此该节点与新闻节点之间的连接强度也高于多数仅发生一次交互的普通用户节点。虽然该节点的 delay_ratio = 1.0，参与时间相对较晚，且 degree_prior = 0.701194 不是样本内最高水平，但其较高的交互强度能够提供更多传播关联信息，因此模型赋予其较高的 trust gate 分数具有一定合理性。该案例表明，信任门控机制能够综合传播交互强度、结构先验和时间延迟等多维信息，对不同节点的信息贡献进行自适应调节，从而增强关键传播信号在图表示学习中的作用。

进一步考察图 3 中 trust gate 分数最低的用户节点，可以发现 u_00272 的原始 gate_score 为 0.497279，是该样本中信任度最低的节点(见表 3)。u_00272 的 delay_ratio = 0.0，说明其较早参与了该事件的传播。

同时，其 degree_prior = 2.713274，在该样本内排名第 1，表明该节点具有较强的结构先验。然而，该节点的 interaction_count = 1，仅发生一次交互，低于高信任节点 u_06753，说明其实际传播参与强度相对有限。虽然该节点在传播时间和结构先验上具有一定优势，但较低的交互强度限制了其所能提供的传播关联信息，因此模型赋予其较低的 trust gate 分数具有一定合理性。该案例表明，信任门控机制能够结合时间位置、结构先验和实际交互强度等因素，对节点的信息贡献进行细粒度调节，从而降低弱交互节点在消息聚合过程中的影响。

Table 3. Attribute comparison between high-trust and low-trust nodes in typical samples
表 3. 典型样本中高/低信任节点属性对比

对比类型	节点 ID	gate_score	interaction_count	degree_prior	delay_ratio
最高信任节点 Top1	u_06753	0.497760	2	0.701194	1.0
最低信任用户节点 Bottom1	u_00272	0.497279	1	2.713274	0.0

上述结果表明，ATEF 能够在具体传播场景中区分不同节点对分类任务的相对贡献。结合表 2 和图 3 可以看出，trust gate 分数并不是由单一的时间位置、结构先验或交互次数决定，而是受到节点交互强度、传播延迟和上下文结构的共同影响。该机制能够在消息传递过程中对不同节点的信息贡献进行平滑调节，从而增强具有较高传播关联信息的节点作用，并降低弱交互节点对图表示学习的干扰。

5.4. 协调扰动攻击下的鲁棒性分析

为进一步考察 ATEF 在极端协同扰动条件下的性能变化，本文设计了基于帖子注入的主攻击实验，并在不同攻击预算下统计模型的检测指标变化。

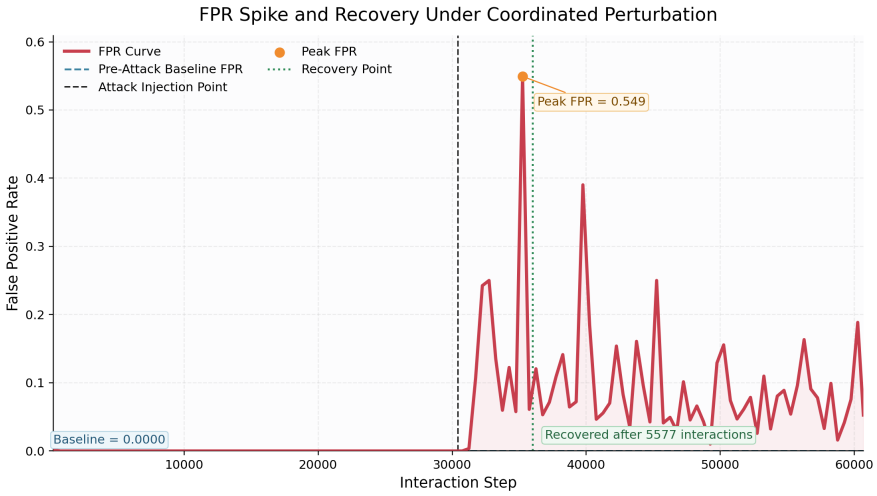

Figure 4. Dynamic variation of ATEF's False Positive Rate (FPR) under coordinated perturbation attacks
图 4. 协调扰动攻击下 ATEF 假阳性率(FPR)的动态变化

为评估 ATEF 在协调扰动场景下的鲁棒性，本文观察了攻击注入后系统假阳性率(FPR)的动态变化过程见图 4。在攻击发生前，系统的 FPR 基本维持在 0 附近，说明模型在正常传播环境下具有较低的误报水平和较稳定的判别边界。当攻击在约 3.0 万交互步附近注入后，FPR 出现明显抬升并迅速进入波动区

间，表明协调扰动会在短时间内显著破坏模型对正常样本的判别稳定性。随后，FPR 在局部达到峰值 0.549，说明在最强冲击阶段，系统误报率一度升高到较高水平。

尽管攻击在短时间内引发了明显的误报冲击，但图中结果表明这种退化并非持续失控。随着后续交互推进，FPR 虽然仍存在一定波动，但整体逐步脱离峰值区间，并在约 5577 次交互后达到恢复点。这表明 ATEF 在面对协调扰动时具备一定的自适应恢复能力，能够在受到显著干扰后重新收敛到较低误报水平。从整体趋势看，该方法对突发扰动并非完全免疫，但其性能退化主要表现为阶段性冲击，而非长期崩溃，这说明所提出框架在复杂传播环境中具有一定的鲁棒性与恢复能力。

6. 消融实验分析

为验证 ATEF 各组成模块的实际贡献，本文进一步开展了主模型消融实验，分别移除 Trust Gate、Temporal Features 以及 Heterogeneous Relations，并在相同实验设置下进行对比。结果见表 4，完整模型 ATEF-full 在测试集上取得了最高的整体性能，其 Accuracy 和 Macro-F1 均为 0.9879，Fake Recall 为 0.9798，AUC (fake)为 0.9992。

Table 4. Ablation results of the main ATEF model
表 4. ATEF 主模型消融实验结果

模型	Accuracy	Macro-F1	Fake Precision	Fake Recall	Fake F1	AUC (fake)
ATEF-full	0.9879	0.9879	0.9960	0.9798	0.9879	0.9992
ATEF w/o Trust Gate	0.9873	0.9873	0.9971	0.9776	0.9873	0.9991
ATEF w/o Temporal Features	0.9486	0.9486	0.9641	0.9328	0.9482	0.9836
ATEF w/o Heterogeneous Relations	0.9856	0.9856	0.9943	0.9770	0.9856	0.9988

当移除 Trust Gate 后，模型的 Macro-F1 由 0.9879 下降至 0.9873，Fake Recall 由 0.9798 下降至 0.9776，说明信任门控机制对模型性能具有一定的正向贡献。尽管该下降幅度相对有限，但其结果表明 Trust Gate 在当前框架下能够对假新闻样本的识别起到补充性作用。当移除 Temporal Features 后，模型性能出现最明显下降，其中 Macro-F1 降至 0.9486，Fake Recall 降至 0.9328，AUC (fake)也由 0.9992 下降至 0.9836。该结果说明，传播过程中的时序信息是 ATEF 中最关键的组成部分之一，能够为谣言检测提供重要判别依据。相比之下，缺少时间特征后，模型对假新闻样本的区分能力明显减弱。当移除 Heterogeneous Relations 后，模型的 Macro-F1 由 0.9879 下降至 0.9856，Fake Recall 由 0.9798 下降至 0.9770。该结果说明，对不同关系类型进行区分建模能够进一步提升模型的表示能力，并对最终检测结果产生稳定增益。

综合来看，三组消融结果共同表明：Temporal Features 是 ATEF 性能提升的主要来源，Heterogeneous Relations 和 Trust Gate 则在此基础上进一步提供了稳定补充，从而共同支撑了完整模型的最优表现。

7. 结论

本文提出了一种面向社交媒体谣言检测的自适应信任评估框架 ATEF。该方法将新闻事件建模为帖子级异构传播图，并通过 trust-gated heterogeneous message passing 对不同传播节点及关系类型的贡献进行自适应调制，从而增强关键传播信号、抑制噪声干扰。实验结果表明，ATEF 在整体分类性能、假新闻召回能力以及复杂传播场景下的稳定性方面均取得了较优表现；与 BiGCN 相比，ATEF 在 Accuracy、Macro-F1 和 Fake Recall 等指标上均表现更优。进一步的攻击实验与消融实验说明，时间特征是模型性能提升的主要来源，异构关系建模与信任门控机制则提供了稳定补充。总体来看，本文验证了将传播结构

建模、时序信息利用与可学习信任调制统一于同一框架中的有效性，也为后续面向更强扰动条件和更复杂真实场景的谣言检测研究提供了参考。

尽管本文提出的 ATEF 在整体分类性能、假新闻召回能力以及复杂传播场景下均取得了较为理想的结果，但仍存在若干不足。首先，当前模型主要基于单一数据集上的帖子级传播图进行验证，其跨数据集、跨平台场景下的泛化能力仍有待进一步考察；其次，本文采用的时间信息仍以相对顺序和伪时间特征为主，对真实时间动态与长程传播依赖的刻画仍不够充分；再次，现有攻击实验主要聚焦于帖子注入场景，对更复杂的结构扰动、后门攻击以及跨阶段协同攻击的适应性仍需进一步研究。未来工作可从更大规模异构社交平台数据的统一建模、更细粒度的时间动态表示学习以及面向强对抗环境的鲁棒优化三个方向展开，以进一步提升模型在真实开放环境中的稳定性、泛化性与实用价值。

参考文献

[1] Vosoughi, S., Roy, D. and Aral, S. (2018) The Spread of True and False News Online. *Science*, **359**, 1146-1151. https://doi.org/10.1126/science.aap9559

[2] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F. and Cha, M. (2016) Detecting Rumors from Microblogs with Recurrent Neural Networks. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (*IJCAI* 2016), New York, 9-15 July 2016, 3818-3824.

[3] Ma, J., Gao, W. and Wong, K. (2017) Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 708-717. https://doi.org/10.18653/v1/p17-1066

[4] Monti, F., Frasca, F., Eynard, D., Mannion, D. and Bronstein, M.M. (2019) Fake News Detection on Social Media Using Geometric Deep Learning. https://arxiv.org/abs/1902.06673

[5] Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., *et al*. (2020) Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 549-556. https://doi.org/10.1609/aaai.v34i01.5393

[6] Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H. (2017) Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, **19**, 22-36. https://doi.org/10.1145/3137597.3137600

[7] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. and Bengio, Y. (2018) Graph Attention Networks. *International Conference on Learning Representations* (*ICLR*), Vancouver, 30 April-3 May 2018.

[8] Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., *et al*. (2019) Heterogeneous Graph Attention Network. *The World Wide Web Conference*, San Francisco, 13-17 May 2019, 2022-2032. https://doi.org/10.1145/3308558.3313562

[9] Zhang, X. and Zitnik, M. (2020) GNNGuard: Defending Graph Neural Networks against Adversarial Attacks. *Advances in Neural Information Processing Systems* 33: *Annual Conference on Neural Information Processing Systems* 2020, *NeurIPS* 2020, 6-12 December 2020, 9263-9275.

[10] Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H. (2020) FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, **8**, 171-188. https://doi.org/10.1089/big.2020.0062

[11] Shu, K., Cui, L., Wang, S., Lee, D. and Liu, H. (2019) dEFEND: Explainable Fake News Detection. *Proceedings of the* 25*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, 4-8 August 2019, 395-405. https://doi.org/10.1145/3292500.3330935

[12] Lu, Y.J. and Li, C.T. (2020) GCAN: Graph-Aware Co-Attention Networks for Explainable Fake News Detection on Social Media. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, 5-10 July 2020, 505-514. https://doi.org/10.18653/v1/2020.acl-main.48

[13] Ying, R., Bourgeois, D., You, J., Zitnik, M. and Leskovec, J. (2019) GNNExplainer: Generating Explanations for Graph Neural Networks. *Advances in Neural Information Processing Systems* 32: *Annual Conference on Neural Information Processing Systems* 2019, *NeurIPS* 2019, Vancouver, 8-14 December 2019, 9240-9251.

[14] Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H. and Zhang, X. (2020) Parameterized Explainer for Graph Neural Network. *Advances in Neural Information Processing Systems* 33: *Annual Conference on Neural Information Processing Systems* 2020, *NeurIPS* 2020, 6-12 December 2020, 19620-19631.