Propagation Structure-Aware Graph Transformer for Robust and Interpretable Fake News Detection

Junyou Zhu
School of Artificial Intelligence, Optics and Electronics
Northwestern Polytechnical University
Xi'an, China
Potsdam Institute for Climate Impact Research
Potsdam, Germany
Junyou.Zhu@pik-potsdam.de

Chao Gao
School of Artificial Intelligence, Optics and Electronics
Northwestern Polytechnical University
Xi'an, China
cgao@nwpu.edu.cn

Ze Yin
Hunan University
College of Computer Science and Electronic Engineering
Changsha, China
zyin@hnu.edu.cn

Xianghua Li*
School of Artificial Intelligence, Optics and Electronics
Northwestern Polytechnical University
Xi'an, China
li_xianghua@nwpu.edu.cn

Juergen Kurths
Potsdam Institute for Climate Impact Research
Potsdam, Germany
Department of Physics
Humboldt-Universität zu Berlin
Berlin, Germany
kurths@pik-potsdam.de

Abstract

The rise of social media has intensified fake news risks, prompting a growing focus on leveraging graph learning methods such as graph neural networks (GNNs) to understand post-spread patterns of news. However, existing methods often produce less robust and interpretable results as they assume that all information within the propagation graph is relevant to the news item, without adequately eliminating noise from engaged users. Furthermore, they inadequately capture intricate patterns inherent in long-sequence dependencies of news propagation due to their use of shallow GNNs aimed at avoiding the over-smoothing issue, consequently diminishing their overall accuracy. In this paper, we address these issues by proposing the **P**ropagation **S**tructure-aware **G**raph **T**ransformer (PSGT). Specifically, to filter out noise from users within propagation graphs, PSGT first designs a noise-reduction self-attention mechanism based on the information bottleneck principle, aiming to minimize or completely remove the noise attention links among task-irrelevant users. Moreover, to capture multi-scale propagation structures while considering long-sequence features, we present a novel relational propagation graph as a position encoding for the graph Transformer, enabling the model to capture both propagation depth and distance relationships of users. Extensive experiments demonstrate the effectiveness, interpretability, and robustness of our PSGT.

CCS Concepts

• **Computing methodologies** → **Knowledge representation and reasoning**; • **Information systems** → **Data mining**.

Keywords

Fake News Detection; Graph Transformer; Social Networks

ACM Reference Format:
Junyou Zhu, Chao Gao, Ze Yin, Xianghua Li, and Juergen Kurths. 2024. Propagation Structure-Aware Graph Transformer for Robust and Interpretable Fake News Detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3672024

1 Introduction

In the current digital age, the rapid growth of the Internet has provided unprecedented opportunities for the production, dissemination, and consumption of fake news [56]. Such deceptive news disrupts public opinion [46] and social harmony [47], but also undermines trust in institutions [13]. Hence, effective fake news detection methods are crucial for public access to reliable information.

Graph Neural Networks (GNNs) [68, 69] have recently demonstrated their efficacy in identifying user propagation patterns, offering vital insights for fake news detection [4, 11, 11, 49, 51]. While these methods have improved detection accuracy, they have not mitigated crucial public concerns regarding the fairness and transparency of such automated systems. This public skepticism has

*Corresponding author.

KDD '24, August 25–29, 2024, Barcelona, Spain.

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3672024

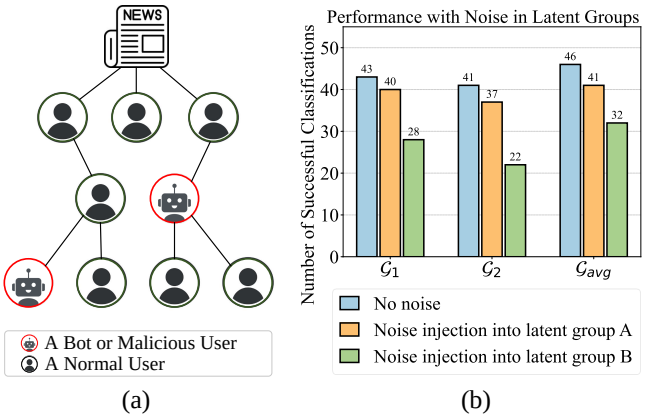


Figure 1: (a) Illustrates noise patterns in news propagation graphs. (b) Empirically observes noise information in two news propagation graphs and the average observation across all graphs in the Politifact dataset [45]. Using k-means [14], users are divided into two groups (A and B) based on their features. Gaussian noise, defined by a Gaussian distribution $\varepsilon \sim N(0, 1)$, is then injected to observe its effects. After 50 runs, successful classifications are counted. Some groups show minimal effects, possibly due to task-irrelevant participants like bots or malicious users.

led many organizations to continue relying on human intervention to address fake news, as it provides a clear explanation for classifying news as either true or false. Thus, enhancing the robustness, transparency, and interpretability of fake news detection models is essential for their final application in real-world scenarios. Nonetheless, enhancing model interpretability to improve user trust remains challenging by two critical issues. **Firstly**, real-world scenarios often involve noise, such as malicious content from users or AI, misleading models to make false predictions and explanations [51]. Ensuring that models can eliminate this noise to provide clear, transparent explanations without sacrificing accuracy is essential for their wider acceptance and practical application. **Secondly**, existing GNN-based detection models, including those designed for interpretability, often struggle to balance capturing news long-sequence propagation structures with avoiding over-smoothing, inherent in their design to learn local patterns [27, 60, 71]. An over-reliance on local propagation structures fails to capture key shifts in news propagation, which can compromise detection accuracy, even if some models provide explanations, contradicting the objectives of practitioners.

Regarding the noisy information in propagation graphs, as illustrated in Figure 1 (a), malicious users are frequently engaged in introducing noise to news comments. Such interactions from adversarial entities often introduce task-irrelevant noise such as disruptive comments, negatively impacting or at best not aiding the model interpretability and prediction [24, 50]. Figure 1 (b) empirically validates this through a case study on news propagation graphs in Political dataset [45]. Intriguingly, specific user groups within propagation graphs, such as group A, display resilience to noise injection, possibly due to the involvement of irrelevant participants such as bots. This resilience implies that group A is primarily composed of noise information, as the presence of substantial task-relevant content would have resulted in a marked decrease in model accuracy following noise injection. In contrast, others, like group B, show pronounced susceptibility, likely because these users featured useful task-relevant information. Such empirical insights provide guidelines for us to enhance both the reliability and accuracy of fake news detection, particularly by mitigating noise impact in news propagation graphs. Despite methods like UPFD [11] use self-attention to focus on key features, recent studies [34, 40, 61] question the efficacy of attention mechanisms in prioritizing impactful features, challenging their utility in enhancing the reliable interpretability in fake news detection.

As for tracking the long-sequence propagation dependencies in news propagation graphs, they remain a critical yet under-explored aspect. These dependencies capture subtle and vital shifts in user engagements, such as comment topic changes and community evolution, during information dissemination. As news diffuses, its cascade depth increases, as evidenced in Appendix D.2 (Figure 6), indicating a substantial propagation distance between some engaged users and the source news. This could explain why existing fake news detection methods [4, 11, 51] based on shallow GNNs often merge features from the original news source into subsequent user nodes. However, such a strategy dilutes a certain task-irrelevant noise signal of users by mixing it with task-relevant information of source news, thereby undermining the capability of models for effective noise filtration, as we demonstrated in Appendix D.2 (Figure 7). While the adoption of vanilla Transformers [54] could intuitively address long-sequence dependencies, recent insights [34, 40, 61] have found that their attention mechanisms do not inherently prioritize features that most significantly impact the output and fail to adequately model the structural relationships within propagation graphs. Although some recent graph-based Transformers have shown being promising in domains like protein prediction [37, 67], their architecture is not well-suited to the information diffusion-based structure characteristic of news propagation graphs, as illustrated in Appendix D.2, restricting their direct applicability in robust and interpretable fake news detection.

In this work, we aim to address the above issues by proposing the **P**ropagation **S**tructure-aware **G**raph **T**ransformer (PSGT), a novel propagation structure-aware graph Transformer designed with innate interpretability and robustness for fake news detection. Building on the information bottleneck (IB) principle [2, 52], PSGT integrates a noise-reduction mechanism within the Transformer architecture, aiming to selectively guide task-relevant information flow across the propagation graph while systematically eliminating extraneous noise. This mechanism initially treats user relationships as a fully connected attention graph, subsequently removing attention links between task-irrelevant graph components to prevent noise transmission. The graph components, such as nodes with abundant connected attention links in the learned noise-reduction attention graph, are eventually identified as task-relevant, and their propagation structure provides interpretability. The model is also expected to be more robust due to its adept noise-filtering capabilities. Furthermore, to model the long-sequence dependencies and the distinctive structural characteristics of news propagation

graphs, we introduce a new positional encoding strategy for the graph Transformer. This strategy models the depth and relational distances among users in the propagation graph, enabling the Transformer architecture to be aware of the propagation structure.

In summary, our contributions are as follows:

- We design a novel noise-reduction mechanism for the graph Transformer based on the IB principle. This mechanism effectively removes task-irrelevant attention links within the self-attention module, yielding interpretable and robust results. To the best of our knowledge, PSGT is the first to propose a unified graph-based IB method for interpretable fake news detection.
- We present a new positional encoding approach rooted in the propagation structure, enabling the model to capture both the propagation structure and long-sequence dependencies, further enhancing both the accuracy and reliability of fake news detection.

2 Related Work

2.1 Fake News Detection

Fake news detection is generally framed as a binary classification problem, aiming to determine the verity or falsity of a specific news article. There are two category methods highly related to our work: **Content-based** and **propagation-network-based** methods. Content-based methods have employed an array of advanced deep learning models such as recurrent neural networks (RNNs) [36] and pre-trained language models to extract semantic signatures from the news articles [26, 64]. In addition to news content, various complementary features have also been explored to enhance prediction accuracy. These supplementary inputs encompass information drawn from knowledge graphs [8, 12, 17], corroborative evidence from external repositories [7, 43], visual cues from associated images [6, 41], and environmental metadata surrounding the news ecosystem [42]. Acknowledging that the spread of fake news is inherently a social phenomenon, propagation-network-based methods incorporate various social attributes into their frameworks [4, 28, 44, 48, 51, 57, 60, 70]. These social indicators include, but are not limited to, user interaction metrics [62], social network connectivity patterns [15, 32, 72], historical posting activity of users [11] and provenance of the news source. Although existing methods have proven being effective in modeling social information, they have overlooked the critical aspect of long-sequence dependencies inherent in news propagation structures. Moreover, these approaches often suffer from low interpretability and increased susceptibility to noise within the propagation graphs, primarily due to their inadequacy in effectively filtering out malicious information.

2.2 Graph Transformer

The Transformer architecture [54], a category of neural networks, has influenced a wide array of tasks, from text to visual data processing [53]. However, pure Transformers lack intrinsic relationships between tokens, thus requiring additional positional encodings for structural understanding. To address this shortcoming, recent studies have explored specialized Transformer adaptations for graph-structured tasks [29, 31]. These adaptations fall into two categories relevant to our study: the first utilizes **absolute positional encodings (APE)** through techniques like Laplacian vectors [23] or random walks [25] to convey nodal structure; the second employs **relative positional encodings (RPE)** to enrich the attention mechanism with structural information derived from graph distances or features generated by GNNs [67]. Additionally, some frameworks incorporate Transformers as modular components in more complex architectures [38]. Despite the progress, these approaches are primarily configured for generic graph-related tasks or specific applications such as recommendation systems [33]. Designing a specialized graph Transformer to tackle fake news detection, particularly by utilizing distinct structural characteristics inherent to news propagation graphs, remains an unsolved challenge.

3 Preliminaries

As preliminaries, we define the problem formulation and concepts.

3.1 Problem Formulation

Fake news detection can be defined as a task of binary classification. The main objective is to train a classifier utilizing a dataset of labeled news and subsequently employ this classifier to determine the veracity of a given test news article.

To provide clarity in our formulation, let us denote our dataset of news articles as $C = \{c_1, c_2, \dots c_m\}$, where c_i denotes the i-th news article and m represents the total count of articles in the dataset. Each news article c_i can be described as a tuple $(y, \mathcal{G})$. Here, y indicates the ground truth label belonging to either F or R, symbolizing whether the news is Fake or Real respectively. Meanwhile, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represents the propagation structure associated with the news article c_i. Specifically, $\mathcal{V} = \{v_0, v_1, \dots v_N\}$, where v_0 represents the source news, and v_i denotes an engaged user of the news. The set of edges $\mathcal{E}$ is defined as $\mathcal{E} = \{e_{ij} \mid A_{ij} \neq 0, \text{for } i, j = 0, \dots, N\}$, where A_{ij} is an element of the adjacency matrix $\mathbf{A}$ that indicates a direct reply or comment relationship from user v_i to user v_j. In this work, we follow prior work [11] to generate the initial features x_i of each node by using pre-trained BERT [10] embeddings to encode the news content and users who comment on the news.

To encapsulate the described framework: Given a news collection C and a set of training labels Y_{train}, our goal is to train a classifier f_ϕ. This function, when presented with unseen test news, should proficiently assign the corresponding veracity labels Y_{test}.

3.2 Transformers on Graphs

3.2.1 Original Transformers. The Transformer architecture is composed of several Transformer layers. Each layer comprises two primary modules: a multi-head self-attention mechanism (MSA) and a feed-forward network (FFN). The head self-attention mechanism is crucial for identifying and comprehending the intrinsic semantic relationships among input tokens. Given an input $\mathbf{X} \in \mathbb{R}^{n \times d}$ for the H head self-attention mechanism, where n is the number of tokens and d denotes the hidden dimension, the mechanism maps $\mathbf{X}$ into three separate spaces: $\mathbf{Q}^h$, $\mathbf{K}^h$, and $\mathbf{V}^h$. More formally, the projections can be described as $\mathbf{Q}^h = \mathbf{X}\mathbf{W}_{\mathbf{Q}}^h$, $\mathbf{K}^h = \mathbf{X}\mathbf{W}_{\mathbf{K}}^h$, and $\mathbf{V}^h = \mathbf{X}\mathbf{W}_{\mathbf{V}}^h$ respectively. For each head h, the self-attention can

then be computed by

$$\text{Attn}^h(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{Q}^h\mathbf{K}^{hT}}{\sqrt{d_H}}\right) \tag{1}$$

where $\mathbf{W}_{\mathbf{Q}}^h \in \mathbb{R}^{d\times d_H}$, $\mathbf{W}_{\mathbf{K}}^h \in \mathbb{R}^{d\times d_H}$, and $\mathbf{W}_{\mathbf{V}}^h \in \mathbb{R}^{d\times d_H}$ are trainable matrices for each head h. H means the total number of heads. d_H denotes the dimension of each head. Subsequently, the output of the self-attention mechanism is combined with a skip-connection, followed by a feed-forward network. Together, these components constitute a Transformer layer, as illustrated below:

$$\mathbf{X}' = \mathbf{X} + \sum_{h=1}^{H} \text{Attn}^h(\mathbf{X})\mathbf{V}^h\mathbf{W}_O^h \tag{2}$$

$$\mathbf{Z} = \text{FFN}(\mathbf{X}') = \text{ReLU}(\mathbf{X}'\mathbf{W}_1)\mathbf{W}_2 \tag{3}$$

where $\mathbf{W}_O^h \in \mathbb{R}^{d_H\times d}$, $\mathbf{W}_1 \in \mathbb{R}^{d\times r}$ and $\mathbf{W}_2 \in \mathbb{R}^{r\times d}$ are trainable matrices.

3.2.2 Graph Transformers. Owing to the inherent equivariance of self-attention to input node permutations, the Transformer consistently produces identical node representations for nodes with the same attributes, regardless of their positions or graph structures. Addressing this, recent studies have introduced positional encoding strategies, i.e., APE and RPE. Our work aligns more closely with RPE, for which we subsequently provide a mathematical formulation. For APE insights, readers are referred to [31]. RPE-based graph Transformer emphasizes relative structural relationships between node pairs. Building on this, most preceding research has adjusted the attention computation described in Equation (1), as follows:

$$\text{Attn}_{RPE}^h(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{Q}^h\mathbf{K}^{hT}}{\sqrt{d_H}} + \mathbf{B}\right) \tag{4}$$

where $\mathbf{B}$ is an $n \times n$ matrix. The entry b_{ij} in $\mathbf{B}$ represents the interaction between nodes v_i and v_j. Different parameterizations of $\mathbf{B}$ yield distinct model architectures. However, prevalent RPE-based graph Transformer methods exhibit two limitations. Firstly, they interpret the network as a fully connected graph within the attention mechanism, lacking the precision to highlight task-relevant features and exclude noise, our method solves these issues, as detailed in Section 4.2. Secondly, these methods, such as GraphGPS [37], are only generic structure-aware, in contrast to our specifically tailored propagation structures-aware RPE strategy, detailed further in Section 4.3.

3.3 Information Bottleneck

Various techniques, including feature selection-based [5] and motif-based methods [65], have been explored to mitigate noise in graphs. However, these methods often impose inherent biased constraints, notably setting constraints on the size and connectivity of task-relevant subgraphs. For instance, in the domain of protein graphs [18], functionally similar group structures exhibit consistency in size. In contrast, domains like news propagation graphs lack this uniformity, with task-relevant substructures displaying variations in size and connectivity.

Informed by the information bottleneck (IB) principle [63], we exclude the task-irrelevant noise by compressing the propagation graph $\mathcal{G}$ without imposing any structural constraints, i.e., solving

$$\max_{\mathcal{G}_s} I\left(\mathcal{G}_s; Y\right), \text{ s.t. } I\left(\mathcal{G}_s; G\right) \leq \gamma, \mathcal{G}_s \in \mathbb{G}_{sub}(\mathcal{G}) \tag{5}$$

where γ is a compression parameter and $\mathbb{G}_{sub}(\mathcal{G})$ represents the set of subgraphs derived from $\mathcal{G}$. $I\left(p;q\right) \triangleq \sum_{p,q}\mathbb{P}(p,q)\log\frac{\mathbb{P}(p,q)}{\mathbb{P}(p)\mathbb{P}(q)}$ represents the mutual information (MI) between variables p and q. Specifically, when handling irregular graph data, the IB framework applies the constraint $I\left(\mathcal{G}_s; G\right) \leq \gamma$. This guides the selection of $\mathcal{G}_s$ to absorb only pivotal information from $\mathcal{G}$ to predict the label Y, thereby maximizing $I\left(\mathcal{G}_s; Y\right)$. Consequently, $\mathcal{G}_s$ plays a crucial role in the interpretability of the model.

Recent studies have also leveraged the IB principle for graph learning [30, 69], primarily focusing on local structures via GNNs. However, as we mentioned before, these methods are not well-suited for fake news detection since news propagation graphs exhibit long-sequence dependencies. To our knowledge, PSGT is the first to utilize graph IB for fake news detection, effectively eliminating noise from propagation graphs while producing interpretable subgraphs that capture both local and long-sequence dependencies.

4 PSGT: Propagation Structure-aware Graph Transformer

In this section, we introduce a novel graph Transformer designed for fake news detection. We begin by providing an overview of the proposed graph Transformer and then present a formal definition of a feasible information bottleneck intended for noise filtration. Next, we detail a unique positional encoding strategy for modeling propagation structures within the graph Transformer. Finally, we discuss the optimization techniques applied to PSGT.

4.1 Overview of the Proposed Graph Transformer

We propose an innovative graph Transformer specifically designed for fake news detection. Notably, this model has the capability to eliminate noise while capturing the long-sequence dependencies characteristic inherent in propagation structures. The main difference between our proposed graph Transformer and the existing graph Transformer architecture lies in the MSA. Within a standard MSA layer, there are H attention heads, each implicitly focusing on different representational subspaces of various nodes. In contrast, our model employs a graph-masking mechanism, compelling these heads to explicitly focus on different subspaces through graph masks.

As illustrated in Figure 2, our approach begins with the integration of a feasible information bottleneck. This mechanism learns a noise-filtered mask graph, $\mathcal{G}_s$ (as detailed in Section 4.2), rooted in a fully connected attention graph. This guides the attention mechanism of graph Transformer to emphasize task-relevant features. To further enhance the ability of the Transformer to capture news propagation patterns, we introduce a new propagation structure modeling strategy. This strategy is tailored for the explicit construction of a structural positional encoding mask, $\mathcal{G}_p$ (as detailed in Section 4.3), grounded in real-world news propagation principles. Based on this, we modify the self-attention score computation

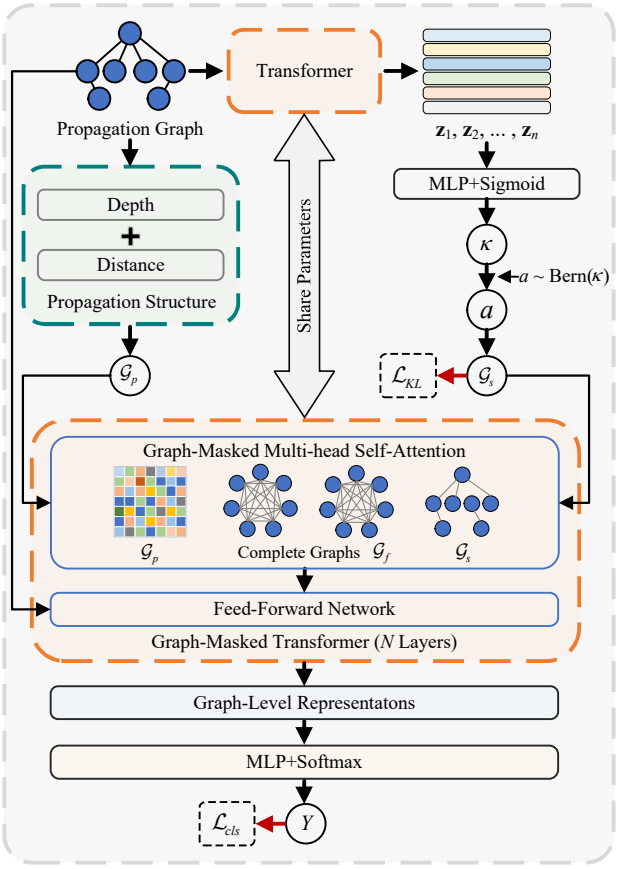


Figure 2: The architecture of PSGT. From a given news propagation graph, we first learn a noise-reduced subgraph $\mathcal{G}_s$ and subsequently construct a more realistic propagation graph $\mathcal{G}_p$. The $\mathcal{G}_s$ primarily retains task-relevant attention links between users for enhanced interpretability, while $\mathcal{G}_p$ captures both the propagation depth and distance relationships among users. Then, using $\mathcal{G}_s$ and $\mathcal{G}_p$ as graph masks, the graph Transformer effectively captures propagation structures and long-sequence dependencies, while filtering out superfluous noise. For label prediction, a binary cross-entropy loss is applied, and a Kullback-Leibler (KL) loss is employed for noise-reduced subgraph learning. Both losses are combined for optimal model training.

described in Equation (4) as:

$$\text{Attn}^h_{Our}(\mathbf{X}) = \text{softmax}\left(\mathcal{M}(\frac{\mathbf{Q}^h\mathbf{K}^{hT}}{\sqrt{d_H}}, \mathbf{A}_i)\right) \tag{6}$$

where $\mathbf{A}_i$ belongs to the set $\{\mathbf{A}_s, \mathbf{A}_p, \mathbf{A}_f\}$. Here, $\mathbf{A}_s$, $\mathbf{A}_p$ and $\mathbf{A}_f = \mathbf{1}_{|\mathcal{V}|\times|\mathcal{V}|}$ represent the adjacency matrices for $\mathcal{G}_s$, $\mathcal{G}_p$, and a fully-connected graph $\mathcal{G}_f$ respectively. The masking function, $\mathcal{M}$, is defined as:

$$\mathcal{M}(x, \lambda) = x + \zeta\lambda \tag{7}$$

where ζ is a sufficiently large value. This approach, while simple, effectively ensures the attention mechanism recognizes structural characteristics and excludes noise information. Given the three types of graph masks $\mathcal{G}_s$, $\mathcal{G}_p$, and $\mathcal{G}_f$, we divide heads into four groups. The first two are masked using $\mathbf{A}_s$ and $\mathbf{A}_p$, whereas the latter two utilize $\mathbf{A}_f$. Note that within $\mathbf{A}_f$, we avoid imposing any structural bias, granting the model the most freedom to learn latent node interrelations. Subsequently, the node representations in l-th Transformer layer, $\mathbf{Z}^l = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$, are obtained by the FFN described in Equation (2).

4.2 A Feasible Graph Information Bottleneck for Noise Filtration

The inherent attention mechanism within the Transformer does not selectively prioritize key features that predominantly influence the output. Consequently, we develop a strategy to learn a noise-filtered graph mask $\mathcal{G}_s$ for the self-attention mechanism in graph Transformer. This aims to enhance the ability of the model to emphasize task-relevant features and simultaneously filter out noise information.

4.2.1 A Feasible Graph IB Objective. Let $\mathbf{A}_h$ denote the self-attention matrix derived from the first head in Transformer layer l via $\text{Attn}^h(\mathbf{X})$. This can be viewed as a fully-connected attention graph, $\mathcal{G}_h = (\mathcal{V}, \mathcal{E}_{full}, \mathbf{Z}^{l-1})$, where $\mathbf{Z}^{l-1}$ is the output of previous Transformer layer and $\mathbf{Z}^0 = \mathbf{X}$. Building on this, PSGT first aims to train an extractor e_Φ with parameter Φ to construct task-relevant attention subgraph $\mathcal{G}_s \in \mathbb{G}_{sub}(\mathcal{G}_h)$. By design, e_Φ encourages the attention mechanism to eliminate task-irrelevant information by removing the noise attention links in $\mathcal{G}_h$, ensuring only task-relevant information is retained in $\mathcal{G}_s$. In essence, $e_\Phi(\mathcal{G}_h)$ offers a distribution over $\mathbb{G}_{sub}(\mathcal{G}_h)$, expressed as $\mathbb{P}_\Phi(\mathcal{G}_s \mid \mathcal{G}_h)$.

Integrating this into Equation (5), we formulate the optimization criterion for e_Φ leveraging the IB principle:

$$\min_{\Phi} -I(\mathcal{G}_s; Y) + \beta I(\mathcal{G}_s; \mathcal{G}_h),\ s.t.\ \mathcal{G}_s \sim e_\Phi(\mathcal{G}_h),\ \beta > 0 \tag{8}$$

We then follow previous works [2, 35, 63] to derive a feasible variational upper bound for terms present in Equation (8). For the term $I(\mathcal{G}_s; Y)$, a parameterized variational approximation $\mathbb{P}_\phi(Y \mid \mathcal{G}_s)$ replaces $\mathbb{P}(Y \mid \mathcal{G}_s)$, yielding the lower bound:

$$I(\mathcal{G}_s; Y) \geq \mathbb{E}_{\mathcal{G}_s, Y}[\log \mathbb{P}_\phi(Y \mid \mathcal{G}_s)] + H(Y) \tag{9}$$

Considering $H(Y)$ is a constant, it is extraneous during optimization. Notably, in our context, $\mathbb{P}_\phi(Y \mid \mathcal{G}_s)$ functions as the predictor or the binary classifier f_ϕ, which predicts the label Y of $\mathcal{G}$ based on $\mathcal{G}_s$. For the other term $I(\mathcal{G}_s; \mathcal{G}_h)$, a variational approximation $\mathbb{Q}(\mathcal{G}_s)$ is introduced for the distribution $\mathbb{P}(\mathcal{G}_s) = \sum_{\mathcal{G}} \mathbb{P}_\Phi(\mathcal{G}_s \mid \mathcal{G}_h)\mathbb{P}_{\mathcal{G}_h}(\mathcal{G}_h)$, resulting in the upper bound:

$$I(\mathcal{G}_s; \mathcal{G}_h) \leq \mathbb{E}_{\mathcal{G}}\left[\text{KL}\left(\mathbb{P}_\Phi(\mathcal{G}_s \mid \mathcal{G}_h) \| \mathbb{Q}(\mathcal{G}_s)\right)\right] \tag{10}$$

Combining these bounds for Equation (8), PSGT objective becomes:

$$\min_{\Phi,\phi} -\mathbb{E}\left[\log \mathbb{P}_\phi(Y \mid \mathcal{G}_s)\right] + \beta\mathbb{E}\left[\text{KL}\left(\mathbb{P}_\Phi(\mathcal{G}_s \mid \mathcal{G}_h) \| \mathbb{Q}(\mathcal{G}_s)\right)\right] \tag{11}$$

In Equation (11), the first term $\mathbb{P}_\phi$ serves as a classifier f_ϕ. Consequently, only the terms $\mathbb{P}_\Phi$ and $\mathbb{Q}$ need further specification, which we will describe in Section 4.2.2 and Section 4.2.3, respectively.

4.2.2 Noise-Reduction Self-Attention Mechanism via $\mathbb{P}_\Phi$. The function $\mathbb{P}_\Phi(\mathcal{G}_s \mid \mathcal{G}_h)$ essentially serves as the extractor e_Φ, designed to extract the task-relevant attention subgraph $\mathcal{G}_s$ from $\mathcal{G}_h$. Notably, due to the densely connected nature of $\mathcal{G}_h$, it contains superfluous and noise attention connections that remain unaddressed by standard self-attention mechanisms.

To address this, we model all attention links as a collection of independent Bernoulli random variables, each parameterized by its respective learned weight κ:

$$\mathbf{A}_s = \bigcup_{u,v \in \mathcal{E}_{full}} \left\{ a_{u,v} \sim \text{Bernoulli}\left(\kappa_{u,v}\right) \right\} \tag{12}$$

The probability κ, governing the sampling of attention links, is optimized within the Transformer architecture. Specifically, the graph $\mathcal{G}_h$ is first encoded by the Transformer, yielding a set of node representations $\{\mathbf{z}_v | v \in \mathcal{V}\}$. Subsequently, for each node pair (u, v), an multi-layer perceptron (MLP) layer combined with a sigmoid activation functions as the extractor e_Φ, transforming the concatenated pair $(\mathbf{z}_u, \mathbf{z}_v)$ into a probability value $\kappa_{u,v} \in [0, 1]$. Here, a lower $\kappa_{u,v}$ indicates a more noise-prone relationship, thus suggesting a reduced weight or exclusion of the attention link.

However, a challenge arises with $\mathbf{A}_s$, given its non-differentiability with respect to κ due to its Bernoulli nature. To solve this, we employ the concrete relaxation technique [19] for the Bernoulli distribution:

$$\text{Bernoulli}\left(\kappa_{u,v}\right) \approx \text{sigmoid}\left(\frac{1}{t}\left(\log\frac{\kappa_{u,v}}{1-\kappa_{u,v}} + \log\frac{\epsilon}{1-\epsilon}\right)\right) \tag{13}$$

where $\epsilon \sim \text{Bernoulli}(0, 1)$ and t acts as the temperature for this concrete distribution. Thus, the distribution of $\mathcal{G}_s$ given $\mathcal{G}_h$, $\mathbb{P}_\Phi(\mathcal{G}_s \mid \mathcal{G}_h)$, can be defined as $\mathbb{P}_\Phi(\mathcal{G}_s \mid \mathcal{G}_h) = \prod_{u,v \in \mathcal{E}_{full}} \mathbb{P}\left(a_{u,v} \mid \kappa_{u,v}\right)$.

4.2.3 Variational Approximation via $\mathbb{Q}$. The bound presented in Equation (10) holds universally for any $\mathbb{Q}(\mathcal{G}_s)$. For optimization efficiency, we also define it as another Bernoulli distribution. Given a graph $\mathcal{G}$ sampled from $\mathbb{P}_\mathcal{G}$, we can get the only $\mathcal{G}_h$ from the first attention head in Transformer. For every node pair (u, v) within the fully connected attention graph $\mathcal{G}_h$, we sample $\tilde{a}_{uv}$ from Bernoulli(ρ), where ρ is a hyperparameter in $[0, 1]$. We subsequently remove all existing attention links in $\mathcal{G}_h$ and adds links (u, v) when $\tilde{a}_{uv}$=1. Suppose the obtained task-relevant attention subgraph is $\mathcal{G}_s$, we have $\mathbb{Q}(\mathcal{G}_s) = \sum_\mathcal{G} \mathbb{P}(\tilde{a} \mid \mathcal{G}_h)\, \mathbb{P}_\mathcal{G}(\mathcal{G}_h)$. Thus, taking into account two Bernoulli distributions, $\mathbb{P}(\mathcal{G}_s \mid \mathcal{G}_h)$ and $\mathbb{Q}(\mathcal{G}_s)$, and omitting constants, the second term of Equation (11) can be described as:

$$\begin{aligned} &\text{KL}\left(\mathbb{P}_\Phi(\mathcal{G}_s \mid \mathcal{G}_h) \,\|\, \mathbb{Q}(\mathcal{G}_s)\right) = \\ &\sum_{(u,v) \in \mathcal{E}_{full}} \kappa_{u,v} \log\frac{\kappa_{u,v}}{\rho} + \left(1-\kappa_{u,v}\right)\log\frac{1-\kappa_{u,v}}{1-\rho} \end{aligned} \tag{14}$$

After extracting subgraph $\mathcal{G}_s$, the classifier f_ϕ employs the masked Transformer, which shares the same parameters as before, to map $\mathcal{G}_s$ into a graph representation using an averaging readout function. This representation is then passed through an MLP layer followed by a softmax function to model the distribution of Y, yielding the distribution $\mathbb{P}_\phi(Y \mid \mathcal{G}_s)$.

4.2.4 Guaranteed Noise Reduction in Attention Graph $\mathcal{G}_h$. PSGT, following the IB principle, enhances graph Transformer's focus on task-relevant features while effectively eliminating superfluous attention links in the self-attention graph $\mathcal{G}_h$. This strategy not only provides model interpretability for the model but also reinforces the robustness. Let $\mathcal{G}_\epsilon$ be a graph solely comprised of those superfluous attention links in $\mathcal{G}_h$. We then give the following theorem, which suggests that minimizing the objective in Equation (8) inherently minimizes the relevance between $\mathcal{G}_s$ and $\mathcal{G}_\epsilon$:

Theorem 4.1. *Let Y be determined solely by the optimal task-relevant attention graph $\tilde{\mathcal{G}}_s$ extracted by e_Φ, and with $\mathcal{G}_\epsilon$ defined as the noise attention graph satisfying $\tilde{\mathcal{G}}_s \cup \mathcal{G}_\epsilon = \mathcal{G}_h$. When we select $\mathcal{G}_s$ equivalent to $\tilde{\mathcal{G}}_s$, it essentially leads to maximizing the objective $I(\mathcal{G}_s; Y) - \beta I(\mathcal{G}_s; \mathcal{G}_h)$.*

Proof of this theorem is detailed in the Appendix A. In essence, Theorem 4.1 suggests that optimizing the objective in Equation (8) forces $\mathcal{G}_s$ to become increasingly unrelated to extraneous attention links in $\mathcal{G}_\epsilon$.

4.3 Propagation Structure Modeling

Up to now, we do not explicitly leverage structural information from observed propagation graphs, which is often recognized as important for modeling propagation behavior. While many existing strategies [4, 51] primarily use adjacency matrices, focusing mainly the first-order propagation relationships among users, our method offers a more realistic view by examining real-world news propagation patterns.

We define the propagation relationship between users in terms of both distance and depth. The distance aspect quantifies the shortest path length between two nodes in the propagation graph, mimicking the fact [3] that if two users are distantly connected in the propagation structure, they might engage in different topics even under the same news, hence sharing less correlation. On the other hand, the depth relationship measures the shortest path from a user to the source news, effectively simulating the prevalent phenomenon of information decay [9] during propagation. As discussions around news deepen, users might veer into topics deviating from the original news theme. These users who participate in such off-topic discussions are less valuable for news classification. Therefore, for any node pair (u, v) in $\mathcal{G}$, our proposed propagation structure modeling module explicitly captures these phenomena, as follows:

$$p_{uv} = e^{\frac{1}{d_{uv}}} \cdot e^{-\frac{d_u + d_v}{2}} \tag{15}$$

In Equation (15), the first and second terms denote the distance and depth relationships, respectively. Here, d_{uv} defines the shortest path length between node u and v, while d_u indicates the shortest path length from node u to the source news. For computational efficiency, we only calculate the shortest path for node pairs whose shortest path is not longer 2, i.e., $d_{uv} \leq 2$. Moreover, by setting the shortest path length to 2, this strategy more intuitively reflects the local structural features of the propagation graph. When combined with the Transformer's ability to characterize long sequence dependencies, indicative of global structure characteristics, the model achieves a more comprehensive understanding of the propagation structures at both local and global levels. This approach yields the graph $\mathcal{G}_p$, with its adjacency matrix $\mathbf{A}_p$ constructed by Equation (15). As we mentioned in Section 4.1, by incorporating $\mathcal{G}_p$ as a graph mask within the graph Transformer, PSGT adeptly captures both the propagation depth and distance relationships among users.

4.4 Optimization and News Classification

The overall optimization objective of PSGT comprises two loss terms: the binary classification loss $\mathcal{L}_{cls}$ defined in Equation (9) and KL loss $\mathcal{L}_{KL}$ defined in Equation (14). For the binary classification loss $\mathcal{L}_{cls}$, we utilize the standard cross-entropy loss. To specify this, we first extract graph-level representations of the news using a readout function:

$$\mathbf{s} = \text{Readout}(\mathbf{A}) = \frac{1}{|\mathcal{V}|} \sum_{v_i \in \mathcal{V}} \mathbf{z}_i \tag{16}$$

Subsequently, this representation is processed through an MLP followed by a softmax function to predict the news label $\hat{Y}$ for a given $\mathbf{s}$ news item c_i. Hence, $\mathcal{L}_{cls}$ is expressed as:

$$\mathcal{L}_{cls}(Y, \hat{Y}) = -(Y \log \hat{Y} + (1 - Y) \log(1 - \hat{Y})) \tag{17}$$

where $Y \in \{0, 1\}$ means the label for each piece of unverified news. Finally, the total optimization objective is then $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{KL}$.

4.5 Comparison with Existing Related Methods

Recent studies applying the IB principle to graph learning have primarily focused on local structures through GNNs [30, 69]. However, in PSGT, the application of IB fundamentally differs from existing approaches, and we have adopted this powerful concept to tackle the specific challenges of interpretable fake news detection. Unlike existing graph IB methods that modify the original graph structure, PSGT leverages the IB principle to filter noisy attention connections within the self-attention layer of Transformers. This is critical because, even if we remove the noise propagation edge between nodes and in the original news propagation graph, the self-attention mechanisms can still naturally and inadvertently rely on toxic correlations between these nodes, giving a higher attention weight to them [40, 61]. Furthermore, GNN-based methods often fall short in capturing the extended sequence dependencies in news propagation, limiting their effectiveness in practical scenarios. PSGT stands out as the first method to employ graph IB for Transformer in fake news detection, successfully filtering noise while generating interpretable subgraphs that comprehensively capture local and long-sequence dependencies.

Additionally, while several graph Transformer methods [31], e.g., GraphGPS [37], have been proposed for applications like protein design, they are not ideally suited for fake news detection, failing to account for unique news propagation features, such as information decay [9]. Our novel graph Transformer model addresses this gap by considering propagation depth and user distance relationships, enhancing both the accuracy and interpretability of fake news detection. Notably, our model excels in using propagation graphs for this purpose. However, the core principles of our model, particularly the propagation structure-based positional encoding strategy, have broader applicability. They can potentially be adapted to various propagation-based graphs, including social, information, and infectious disease networks [58]. We aim to explore these applications in future work.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets. Our study focuses on the impact of news propagation structure and user engagement in fake news detection, utilizing the FakeNewsNet dataset, which includes PolitiFact and GossipCop sub-datasets. Each entry comprises a news piece, related Twitter posts, user interactions, and a "Real" or "Fake" label assigned by experts. We preprocess the data following the methods in [11], using 75% of the articles for training and the remainder for testing. Detailed dataset statistics are available in Appendix C.1.

5.1.2 Comparison Methods. To explore the impact of news propagation structures, we benchmark our PSGT algorithm[1] against various known methods, which span both content-based methods (**M1**) and propagation-structure-based methods (**M2**). In the content-based category, we consider **CSI** [39], which employs an LSTM network to learn sequential retweet features, as well as **Bert-MLP** and **Spacy-MLP**, which classify news using MLPs based on 768-dimensional and 300-dimensional features encoded by pretrained BERT [10] and spaCy [16] word2vec models, respectively. We also consider **UniPF** [59], which develops a semantic-clustering strategy for fake news detection. For propagation-network-based methods, we include **GCN** [22], a well-known graph neural network (GNN) that uses a message propagation mechanism to learn representations for users and news items; **GAT** [55], which distinguishes node importance through an attention mechanism; **GraphGPS** [37], a graph Transformer approach incorporating positional encoding; **BiGCN** [4], which employs two separate GCNs for modeling both the propagation-directed and dispersion-directed graphs; **UPFD** [11], which integrates the social history of users and news features into a GNN architecture; **GACL** [51], introducing an adversarial contrastive learning strategy into rumor detection; **EBGCN** [60], a probabilistic model that captures propagation uncertainty by encoding propagation trees with edge-enhanced Bayesian networks; and **GSAT** [30], which employs a graph information bottleneck technique for optimization purposes; **DECOR** [62], which refines social graph structures by applying a degree-correction mechanism to better align with real-world social dynamics; **HGFND** [20], which uses hypergraph neural networks to capture complex relational data in news propagation networks, enhancing the detection of fake news through higher-order relationship modeling; and **FinerFact** [21], which employs a fine-grained reasoning framework and dual-channel graph network for subtle evidence differentiation, enhancing detection capabilities and explainability.

Detailed information on evaluation metrics, experimental settings, and implementation is provided in Appendix C.

5.2 Overall Performance

Tables 1 and 2 demonstrate that PSGT consistently outperforms competing baseline methods across both datasets. Several key observations can be informed from these results: **(1) Advantage of Propagation-Structure-Based Methods:** Methods based on propagation structure yield a marked improvement in performance over those relying solely on content. This aligns well with expectations,

[1]The implementation code is available at https://github.com/JYZHU03/PSGT.

Table 1: Comparison of PSGT and baselines (M1: Content-based, M2: Propagation-structure-based) on PolitiFact.

	Methods	ACC	Pre	Rec	F1
M1	CSI	0.734	0.672	0.550	0.688
	Bert-MLP	0.853	0.892	0.833	0.825
	SpaCy-MLP	0.375	0.124	0.136	0.272
	UniPF	0.760	0.783	0.800	0.754
M2	GCN	0.833	0.958	0.766	0.831
	GAT	0.895	0.930	0.900	0.890
	GraphGPS	0.903	0.900	0.818	0.857
	BiGCN	0.854	0.960	0.800	0.851
	UPFD	0.875	0.900	0.900	0.866
	GACL	0.875	0.961	0.833	0.871
	EBGCN	0.896	0.898	0.909	0.891
	GSAT	0.854	0.947	0.750	0.852
	DECOR	0.907	0.951	0.911	0.918
	HGFND	0.911	0.947	0.907	0.911
	FinerFact	0.909	0.919	0.904	0.917
Our	**PSGT**	**0.917**	**0.966**	**0.918**	**0.922**

Table 2: Comparison of PSGT and baselines (M1: Content-based, M2: Propagation-structure-based) on GossipCop.

	Methods	ACC	Pre	Rec	F1
M1	CSI	0.866	0.892	0.840	0.866
	Bert-MLP	0.962	0.968	0.954	0.961
	SpaCy-MLP	0.501	0.498	0.534	0.333
	UniPF	0.933	0.932	0.933	0.932
M2	GCN	0.957	0.931	0.984	0.957
	GAT	0.961	0.945	0.971	0.960
	GraphGPS	0.968	0.972	0.952	0.960
	BiGCN	0.951	0.923	0.982	0.953
	UPFD	0.965	0.972	0.957	0.965
	GACL	0.976	0.971	0.972	0.976
	EBGCN	0.964	0.966	0.962	0.963
	GSAT	0.956	0.949	0.962	0.955
	DECOR	0.972	0.962	0.961	0.956
	HGFND	0.974	0.963	0.973	0.974
	FinerFact	0.832	0.862	0.878	0.869
Our	**PSGT**	**0.980**	**0.974**	**0.987**	**0.980**

as the former captures both the propagation structure and the features of news into unified node representations, thereby amplifying accuracy. **(2) Superiority of PSGT:** Among the techniques that leverage propagation structure, our proposed PSGT outshines all competitors. This highlights the effectiveness of PSGT to extract both structural patterns and feature information. **(3) Comparison with Other GNN-Based Methods:** Specifically, PSGT outperforms GNN-based methods like BiGCN, UPFD, GACL, and EBGCN, which are developed with the aim of detecting fake news. Their underperformance can be attributed to their reliance on shallow GNN architectures that are not capable of capturing long-sequence features. **(4) Comparison with Graph Transformer and IB Methods:** Interestingly, even methods such as GraphGPS, which accounts for long-sequence features, and GSAT, which employs an information bottleneck, do not surpass PSGT. This likely stems from their inability to simultaneously capture the unique graph structure inherent to news propagation while filtering out extraneous noise. These insights collectively validate the effectiveness of PSGT, which thoughtfully considers both propagation structure and noise filtration, in detecting fake news on real-world social media.

5.3 Robustness

In real-world fake news tasks, robustness is essential, especially since fake news propagation graphs often face noise disruptions. To test the robustness of our model, we incorporated Gaussian noise, characterized by a Gaussian distribution $\varepsilon \sim N(0, 1)$, into node features and removed 20% of the edges. When assessing accuracy by introducing varying noise levels to node features, our results in Figure 3 revealed that our PSGT consistently outperformed baselines, even when some, like GAST, employed strategies like IB for key feature selection, or GACL used adversarial learning to eliminate noise. Intriguingly, we found that a certain proportion of noise could enhance performance. This finding aligns with previous research [4] indicating a minimal influence from non-root nodes in news propagation graphs, with optimal noise levels refocusing the model on the root node, thereby improving its overall performance.

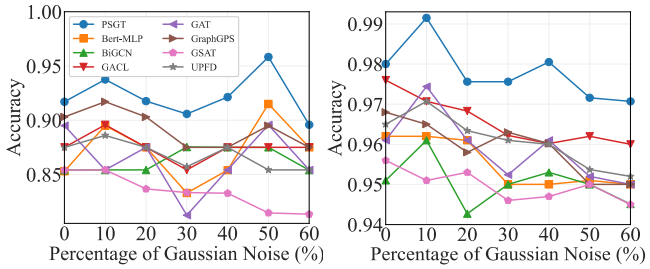


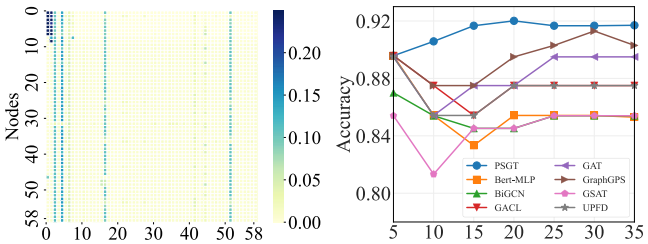
Figure 3: RobustnessRobustness test on Politifact and Gossipcop. The x-axis represents the percentage of Gaussian noise injected, while the y-axis denotes accuracy. PSGT consistently surpasses baselines across various noise levels.

5.4 Interpretability

An interpretable result aids individuals in understanding which parts of the data the model leverages, with minimal cognitive effort [66]. When the model exhibits high accuracy, its generated explanations do not only clarify the model's workings but also shed light on the dataset. For instance, they can highlight which users or propagation structures are pivotal for fake news detection. We begin by visualizing a propagation graph from PolitiFact comprising 58 nodes. This graph size is optimal for visualization, and

other graphs display similar patterns. As shown in Figure 4 (a), the source news node (*id*=0) and its immediate neighbors carry higher attention weights, indicating the emphasis of the model on early propagation. This aligns with prior research [32], highlighting that initial engaged users often provide pertinent comments.

Building on the observation that the most interpretable aspects often reside in the initial propagation structure, an intriguing question arises: for a user with a limited cognitive resources, how effectively can they derive insights? To address this, we execute an automatic quantitative experiment, wherein we extract small early propagation subgraphs from PolitiFact at various cognitive loads. Specifically, beginning with the root node, we employ a breadth-first search algorithm to extract 5% to 35% of the nodes and their associated edges. These subsets represent varying cognitive loads. These subgraphs are then input into our models. Figure 4 (b) demonstrates that PSGT surpasses the baselines across all cognitive loads and performs exceptionally well, even at minimal cognitive demands. Interestingly, the baselines also perform well at a mere 5% cognitive load. This can be attributed to the limited noise in the 5% data subset, allowing these models to primarily depend on the source news without disruptive interactions. However, in contrast to the baselines, our PSGT approach incorporates a noise-filtering mechanism. Hence, as cognitive load increases, the model can still extract value without being overly affected by noise.



(a) Attention Weights (b) Cognitive Budget

Figure 4: Interpretability results from PolitiFact data. (a) Heatmap of attention weights, highlighting the crucial part of early propagation structure surrounding the source news node (*id*=0). (b) PSGT and baseline performance across cognitive loads, demonstrating PSGT can achieve competitive performance even at minimal cognitive demands.

6 Conclusion

In this paper, we have introduced PSGT, a propagation structure-aware graph Transformer, specifically designed for interpretable fake news detection. By incorporating a graph Information Bottleneck strategy, PSGT inherently provides robust and interpretable results. Guided by the IB principle, our proposed graph Transformer effectively filters out task-irrelevant information through a noise-reduction attention graph, enhancing the reliability in interpretability. Additionally, we have presented a unique positional encoding strategy tailored for the Transformer architecture, enabling PSGT to capture both long-sequence dependencies and the propagation structure effectively. Extensive evaluations comparing PSGT with various known methods, including those based on GNNs, graph Transformers, and IB principles, demonstrate the effectiveness of PSGT in providing robust and interpretable fake news detection outcomes.

7 Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62271411, U22A2098, 62261136549 and 11931015), the China Scholarship Council scholarship.

References

[1] Alessandro Achille and Stefano Soatto. 2018. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research* 19, 1 (2018), 1947–1980.
[2] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *Proceedings of the 5th International Conference on Learning Representations*. OpenReview.net.
[3] Venkatesh Bala and Sanjeev Goyal. 2000. A noncooperative model of network formation. *Econometrica* 68, 5 (2000), 1181–1229.
[4] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with Bi-directional graph convolutional Networks. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press, 549–556.
[5] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, 882–891.
[6] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Tun Lu, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*. ACM, 2897–2905.
[7] Zhendong Chen, Siu Cheung Hui, Fuzhen Zhuang, Lejian Liao, Fei Li, Meihuizi Jia, and Jiaqi Li. 2022. EvidenceNet: Evidence fusion network for fact verification. In *Proceedings of the ACM Web Conference 2022*. ACM, 2636–2645.
[8] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. DETERRENT: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 492–502.
[9] Daryl J Daley and David G Kendall. 1964. Epidemics and rumours. *Nature* 204, 4963 (1964), 1118–1118.
[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. Association for Computational Linguistics, 4171–4186.
[11] Yingtong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2051–2055.
[12] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. KAN: Knowledge-aware attention network for fake news detection. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. AAAI Press, 81–89.
[13] Shubham Gupta, Narendra Yadav, Suman Kundu, and Sainathreddy Sankepally. 2023. FakEDAMR: Fake News Detection Using Abstract Meaning Representation Network. In *International Conference on Complex Networks and Their Applications*. Springer, 308–319.
[14] Greg Hamerly and Charles Elkan. 2003. Learning the k in k-means. In *Advances in Neural Information Processing Systems*. MIT Press, 281–288.
[15] Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor detection on social media with event augmentations. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2020–2024.
[16] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear* 7, 1 (2017), 411–420.
[17] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to The knowledge: graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 754–763.
[18] Ylva Ivarsson and Per Jemth. 2019. Affinity and specificity of motif-based protein–protein interactions. *Current opinion in structural biology* 54 (2019), 26–33.
[19] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *Proceedings of the 5th International Conference on Learning Representations*. OpenReview.net.

[20] Ujun Jeong, Kaize Ding, Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2022. Nothing stands alone: Relational fake news detection with hypergraph neural networks. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 596–605.

[21] Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5746–5754.

[22] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations*. OpenReview.net.

[23] Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. 2021. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems*. 21618–21629.

[24] Thai Le, Suhang Wang, and Dongwon Lee. 2020. MALCOM: Generating malicious comments to attack neural fake news detection models. In *Proceedings of the 20th IEEE International Conference on Data Mining*. IEEE, 282–291.

[25] Pan Li, Yanbang Wang, Hongwei Wang, and Jure Leskovec. 2020. Distance encoding: Design provably more powerful neural networks for graph representation learning. In *Advances in Neural Information Processing Systems*.

[26] Qifei Li and Wangchunshu Zhou. 2020. Connecting the dots between fact verification and fake news detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. 1820–1825.

[27] Leyuan Liu, Junyi Chen, Zhangtao Cheng, Wenxin Tai, and Fan Zhou. 2023. Towards Trustworthy Rumor Detection with Interpretable Graph Structural Learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4089–4093.

[28] Guanghui Ma, Chunming Hu, Ling Ge, Junfan Chen, Hong Zhang, and Richong Zhang. 2022. Towards robust false information detection on social networks with contrastive learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1441–1450.

[29] Liheng Ma, Chen Lin, Derek Lim, Adriana Romero-Soriano, Puneet K. Dokania, Mark Coates, Philip H. S. Torr, and Ser-Nam Lim. 2023. Graph inductive biases in transformers without message passing. In *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202. PMLR, 23321–23337.

[30] Siqi Miao, Mia Liu, and Pan Li. 2022. Interpretable and generalizable graph learning via stochastic attention Mechanism. In *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162. PMLR, 15524–15543.

[31] Erxue Min, Runfa Chen, Yatao Bian, Tingyang Xu, Kangfei Zhao, Wenbing Huang, Peilin Zhao, Junzhou Huang, Sophia Ananiadou, and Yu Rong. 2022. Transformer for graphs: An overview from architecture perspective. *arXiv preprint arXiv:2202.08455* (2022).

[32] Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-conquer: Post-user interaction network for fake news detection on Social Media. In *Proceedings of the ACM Web Conference 2022*. ACM, 1148–1158.

[33] Erxue Min, Yu Rong, Tingyang Xu, Yatao Bian, Da Luo, Kangyi Lin, Junzhou Huang, Sophia Ananiadou, and Peilin Zhao. 2022. Neighbour interaction based click-through rate prediction via graph-masked transformer. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 353–362.

[34] Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4206–4216.

[35] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. PMLR, 5171–5180.

[36] Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press, 490–497.

[37] Ladislav Rampásek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. 2022. Recipe for a general, powerful, scalable graph transformer. In *Proceedings of the Advances in Neural Information Processing Systems*. 14501–14515.

[38] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*.

[39] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 797–806.

[40] Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable?. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics, 2931–2951.

[41] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022. A duo-generative approach to explainable multimodal COVID-19 misinformation Detection. In *Proceedings of the ACM Web Conference 2022*. ACM, 3623–3631.

[42] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 4543–4556.

[43] Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating pattern- and fact-based fake news detection via model preference learning. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*. ACM, 1640–1650.

[44] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 395–405.

[45] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8, 3 (2020), 171–188.

[46] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[47] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. ACM, 312–320.

[48] Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management* 58, 5 (2021), 102618.

[49] Xing Su, Jian Yang, Jia Wu, and Yuchen Zhang. 2023. Mining user-aware multi-relations for fake news detection in large scale online social networks. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*. ACM, 51–59.

[50] Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and Philip S. Yu. 2022. Graph structure learning with variational information bottleneck. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. 4165–4174.

[51] Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor detection on social media with graph adversarial contrastive Learning. In *Proceedings of the ACM Web Conference 2022*. ACM, 2789–2797.

[52] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *Proceedings of the 2015 IEEE Information Theory Workshop*. IEEE, 1–5.

[53] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. 2021. Going deeper with image transformers. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*. IEEE, 32–42.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[55] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*. OpenReview.net.

[56] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[57] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 849–857.

[58] Zhen Wang, Dongpeng Hou, Chao Gao, Xiaoyu Li, and Xuelong Li. 2023. Lightweight source localization for large-scale social networks. In *Proceedings of the ACM Web Conference 2023*. 286–294.

[59] Lingwei Wei, Dou Hu, Yantong Lai, Wei Zhou, and Songlin Hu. 2022. A unified propagation forest-based framework for fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*. 2769–2779.

[60] Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. 2021. Towards propagation uncertainty: Edge-enhanced bayesian graph convolutional Networks for Rumor Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 3845–3854.

[61] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 11–20.

[62] Jiaying Wu and Bryan Hooi. 2023. DECOR: Degree-corrected social graph refinement for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2582–2593.

[63] Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020. Graph information bottleneck. In *Advances in Neural Information Processing Systems*.

[64] Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*. ACM, 2501–2510.

[65] Carl Yang, Mengxiong Liu, Vincent W. Zheng, and Jiawei Han. 2018. Node, motif and subgraph: Leveraging network functional blocks through structural convolution. In *Proceedings of the IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining*. IEEE Computer Society, 47–52.

[66] Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement subgraph reasoning for fake news detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2253–2262.
[67] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Yanming Shen, and Tieyan Liu. 2021. Do transformers really perform badly for graph representation?. In *Advances in Neural Information Processing Systems*. 28877–28888.
[68] Junchi Yu, Jie Cao, and Ran He. 2022. Improving subgraph recognition with variational graph information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 19374–19383.
[69] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. 2021. Graph information bottleneck for subgraph recognition. In *Proceedings of the 9th International Conference on Learning Representations*. OpenReview.net.
[70] Kaiwei Zhang, Junchi Yu, Haichao Shi, Jian Liang, and Xiao-Yu Zhang. 2023. Rumor detection with diverse counterfactual evidence. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3321–3331.
[71] Junyou Zhu, Xianghua Li, Chao Gao, Zhen Wang, and Jurgen Kurths. 2021. Unsupervised community detection in attributed networks based on mutual information maximization. *New Journal of Physics* 23, 11 (2021), 113016.
[72] Junyou Zhu, Chunyu Wang, Chao Gao, Fan Zhang, Zhen Wang, and Xuelong Li. 2021. Community detection in graph: an embedding method. *IEEE Transactions on Network Science and Engineering* 9, 2 (2021), 689–702.

A Proof of Theorem 4.1

Proof. We adopt similar technicalities in [1] to prove Theorem 4.1. Consider a fully connected attention graph, denoted as $\mathcal{G}_h$, which is determined by noise attention links $\mathcal{G}_\epsilon$ and the authenticity label Y of the news c. Let the optimal task-relevant attention graph, $\mathcal{G}_s$, depend on $\mathcal{G}_\epsilon$ solely via $\mathcal{G}_h$. This relationship can be captured by the Markov Chain $\langle (Y, \mathcal{G}_\epsilon) \rightarrow \mathcal{G}_h \rightarrow \mathcal{G}_s \rangle$. Given that $\mathcal{G}_\epsilon$ represents noise attention links, and is therefore unrelated and independent of Y, it follows that $H(Y \mid \mathcal{G}_\epsilon) = H(Y)$ and $H(Y \mid \mathcal{G}_\epsilon; \mathcal{G}_s) \leq H(Y \mid \mathcal{G}_s)$. Invoking the data processing inequality, we deduce:

$$\begin{aligned} I(\mathcal{G}_s; \mathcal{G}_h) &\geq I(\mathcal{G}_s; Y, \mathcal{G}_\epsilon) \\ &= I(\mathcal{G}_s; \mathcal{G}_\epsilon) + I(\mathcal{G}_s; Y \mid \mathcal{G}_\epsilon) \\ &= I(\mathcal{G}_s; \mathcal{G}_\epsilon) + H(Y \mid \mathcal{G}_\epsilon) - H(Y \mid \mathcal{G}_\epsilon; \mathcal{G}_s) \\ &\geq I(\mathcal{G}_s; \mathcal{G}_\epsilon) + H(Y) - H(Y \mid \mathcal{G}_s) \\ &= I(\mathcal{G}_s; \mathcal{G}_\epsilon) + I(\mathcal{G}_s; Y) \end{aligned} \tag{18}$$

This concludes the proof. □

B Complexity Analysis

The time complexity of PSGT is primarily influenced by the construction of the propagation graph and the design of the Transformer architecture. Specifically, we employ the Dijkstra algorithm to determine the depth relationships between the source news and all users. This has a time complexity of $O(|\mathcal{V}| \log |\mathcal{V}| + |\mathcal{E}| \log |\mathcal{V}|)$. Retrieving second-order neighbors for all nodes incurs a cost of $O(|\mathcal{E}||\mathcal{V}|)$. Given that our Transformer architecture involves a calculation of attention similar to standard Transformer designs, its complexity stands at $O(|\mathcal{V}|^2)$. Regarding the IB principle, the first term has a complexity of $O(|\mathcal{V}|)$, while the second term has a complexity of $O(d|\mathcal{V}|)$. Consequently, for a training dataset comprising m news articles, the overall time complexity of PSGT amounts to $O(md|\mathcal{V}|^2)$.

C Implementation Details

C.1 Dataset and Graph Statistics

We present a detailed statistical analysis of the datasets utilized in this study, as shown in Table 3. The dataset used has few nodes per graph, so the average depth of news propagation graphs does not exceed 4, though some reach higher depths, as Figure 6 illustrates.

C.2 Experimental Settings and Implementation

Following the prior work [11], we initialize word embeddings using 768-dimensional feature vectors that have been pre-trained via BERT on extensive corpora. These settings are applied to both the PolitiFact and GossipCop datasets. We stack two Transformer layers with 128 hidden dimensions each. Training is conducted over 600 epochs, with model parameters updated using the Adam optimizer at a learning rate of 0.001. β and ζ are not tuned and are set to 1 and 10^3, respectively. Temperature parameter t used in Equation (13) is also not tuned, and we set it as 1 for all datasets. The number of Transformer layers is set to 2 for all datasets. To ensure the robustness of our findings, we average the experimental outcomes across ten independent runs. Experiments are performed using the PyTorch framework on a Linux server equipped with Nvidia V100 GPUs.

C.3 Evaluation Metrics

In line with previous works [11, 51], we employ a quartet of commonly utilized evaluation metrics to assess the effectiveness of various fake news detection techniques. These metrics include Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and F1 Score (F1).

D More Experiment Results

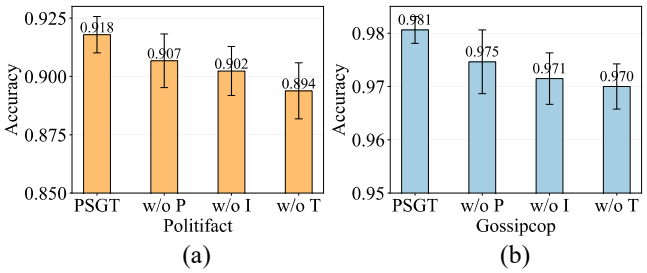


Figure 5: Ablation study on Politifact and Gossipcop datasets. Bars represent model accuracy for configurations: full (PSGT), without news propagation structure (w/o P), without information bottleneck (w/o I), and with GAT instead of Transformer (w/o T). Error bars denote standard deviations.

D.1 Ablation Study

Our model comprises three essential modules: the News Propagation Structure, the Information Bottleneck (IB) Principle, and the Transformer Architecture. These modules are responsible for simulating propagation behavior, eliminating irrelevant information, and capturing long-range dependencies, respectively. To quantify the contributions of each of these modules, we conducted an ablation study. For this study, we created three model variants:

- **"w/o N"** removes the News Propagation Structure module.
- **"w/o I"** is a variant without the IB Principle

Table 3: Statistic of the experimental datasets

Dataset	Graphs (Fake)	Total Nodes	Total Edges	Avg. Nodes per Graph
Politifact	314 (157)	41,054	40,740	131
Gossipcop	5464 (2732)	314,262	308,798	58

- **"w/o T"** replaces the Transformer Architecture with the Graph Attention Networks (GAT) architecture, which is a type of GNNs that also has a self-attention mechanism.

Results presented in Figure 5 clearly show that the omission of any of these modules results in a decrease in model performance. This leads us to several key observations: (1) Both news features and propagation structures are essential for effective model learning. (2) The IB Principle contributes to the model by effectively filtering out noisy data, thereby enhancing accuracy. (3) The Transformer Architecture outperforms GAT, despite both having self-attention mechanisms, emphasizing the importance of long-sequence-dependent features in news propagation graphs. For an extended comparison between GNNs and Transformer architectures, please refer to Appendix D.2 (Figure 7).

D.2 Comparing PSGT and GNN-based Fake News Detection with News Root Concatenation

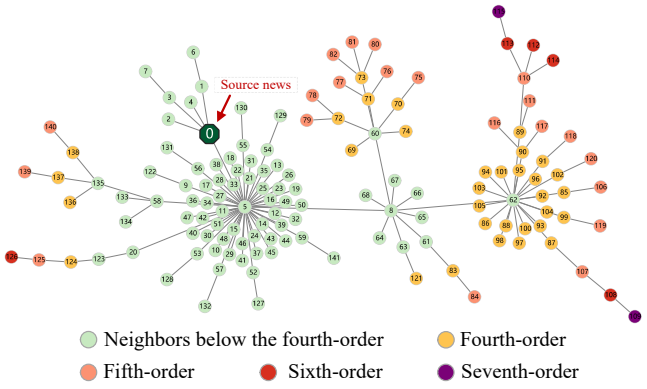


Figure 6: Visualization of a 141-node, information diffusion-based news propagation graph from Politifact. The source news node is labeled with *id*=0, with varying node colors representing the different orders of propagation distance from the source news node. Numerous nodes exhibit a propagation distance exceeding three orders from the source news node.

In realistic information diffusion-based news propagation graphs, the depth of the propagation cascade can be substantial, indicating a large propagation distance between some engaged users and the source news. For instance, in a news propagation graph with 141 nodes as depicted in Figure 6, many nodes have a propagation distance from the source news exceeding three orders. When employing shallow GNNs-based fake news detection methods in such

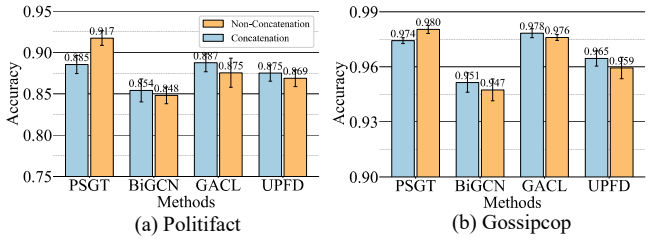


Figure 7: Performance comparison of several GNN-based fake news detection methods in concatenation and non-concatenation scenarios. Unlike existing GNN-based fake news detection methods which benefit from concatenation, PSGT demonstrates superior performance without concatenation, underlining its efficacy in handling long-sequence dependencies in the propagation graph.

scenarios, the information from these nodes cannot be efficiently passed to the source news, nor can it capture the source news information adequately. This limitation explains why existing fake news detection techniques relying on shallow GNNs often incorporate features from the source news into subsequent user nodes.

However, this strategy has a disadvantage as it dilutes the noise signal of user nodes by blending it with source news, thereby harming the ability of models for effective noise filtration. For example, when concatenating the features of source news to a noisy user node, the noise information gets diluted by the relevant features of the source news, which negatively impacts model learning. Unlike existing GNN-based fake news detection methods, our proposed PSGT model has an inherent ability to capture long sequence dependencies without the need for extra concatenation. This capability allows non-noise but distant nodes to self-adapt and capture information from the source news efficiently, without the risk of introducing noise to the source news. To empirically validate this, we conducted an experiment comparing the performance between our PSGT and GNN-based fake news detection methods, with and without concatenation of source news features.

Figure 7 shows typical GNN-based methods improve performance when concatenating source news features to user nodes, whereas our PSGT exhibits the opposite behavior. This finding highlights the inefficacy of shallow GNNs in capturing long-sequence features within the propagation graph effectively. As for the observed decline in PSGT performance, a reasonable explanation is that after injecting all nodes with root node information, the noise node information is diluted by the features of task-relevant news nodes, leading the model to misidentify these noise nodes as non-noise nodes. Consequently, their original noise information is transmitted to other nodes, deteriorating the quality of the learned node representation and thereby reducing experimental accuracy.