HGAT: Hierarchical Graph Attention Network for Fake News Detection

Yuxiang Ren,¹ Jiawei Zhang¹ ¹IFM lab, Florida State University yuxiang@ifmlab.org, jiawei@ifmlab.org

Abstract

The explosive growth of fake news has eroded the credibility of medias and governments. Fake news detection has become an urgent task. News articles along with other related components like news creators and news subjects can be modeled as a heterogeneous information network (HIN for short). In this paper, we focus on studying the HINbased fake news detection problem. We propose a novel fake news detection framework, namely Hierarchical Graph Attention Network (HGAT) which employs a novel hierarchical attention mechanism to detect fake news by classifying news article nodes in the HIN. This method can effectively learn information from different types of related nodes through the node-level and schemalevel attention. Experiments with real-world fake news data show that our model can outperform text-based models and other network-based models. Besides, the experiments also demonstrate the expandability and potential of HGAT for heterogeneous graphs representation learning in the future.

1 Introduction

With the explosive growth, fake news has already caused serious threats to the public's factual judgment and the credibility of governments. Especially with the wide use of social platforms, they facilitate the generation and dissemination of fake news. For example, during the 2016 US presidential election, a lot of fake news about presidential candidates is spread on various social platforms as well [Jin et al., 2017], e.g., 115 pro-Trump fake stories that were shared on Facebook a total of 30 million times, and 41 pro-Clinton fake stories shared a total of 7.6 million times are observed in [Allcott and Gentzkow, 2017]. Such a huge amount of widely spread fake news have greatly destroyed the public persona of candidates and misled the judgment of voters. It has become very critical to detect fake news on social media in time to block the spread and refute them.

There are significant differences between fake news and traditional fraudulent information. First, fake news is usually edited by the creators to achieve the purpose of misleading

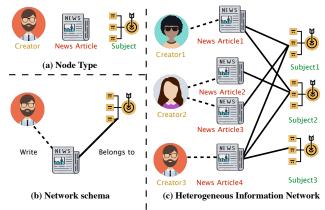


Figure 1: An illustrative example of a heterogenous information network based on PoliticFact data (News-HIN). (a) Three types of nodes (i.e., Creator, News article, Subject). (b) Network schema of News-HIN (c) A News-HIN consists three types of nodes and two types of links.

the public. For instance, news about the same event published by different creators are highly similar in most contents, but fake news carries some fake contents among the objective statements. Although the proportion of these fake contents is negligible, this is enough to make the news as a harmful fake one. Second, for traditional suspicious information, like spam [Akoglu et al., 2013; Xie et al., 2012a; Xie et al., 2012b], people instinctively have a precautionary mentality that makes themselves less likely to be deceived. But for news, people usually actively search, receive, and share without being on guard about authenticity. Third, spams are normally easier to be detected because of the abundant regular messages; yet, detecting fake news is incredibly challenging, since news is very time-sensitive. The evidencecollecting about news in the past can not benefit the detection of emerging fake news apparently.

These characteristics of fake news make the detection more challenging. In order to detect fake news more effectively, it's necessary to mine meaningful information from different views instead of focusing on the news contents solely. In fact, fake news does not exist independently in the form of an article, like news creators and news subjects relating to news articles also exist in online social media. The information from

news creators and news subjects describe the news in a more comprehensive view and help us more thoroughly eliminate fake news and relating components. In detail, for the news creators, we are able to collect the profile information and other supplementary knowledge. For the news subjects, the background knowledge and other related information can be collected to support the news detection. News articles along with other related components can be modeled as a heterogeneous information network (HIN for short) [Sun and Han, 2012; Shi *et al.*, 2017]. HINs have a powerful capability of representing rich information, and we formulate fake news detection as the node classification problem in the HIN in this paper. We present an illustrative example of a news oriented heterogenous information network (News-HIN) in Figure 1.

Problem Studied: In this paper, we propose to study the HIN-based fake news detection problem. We aim at identifying fake news articles in the HIN with the support of various types of heterogeneous information sources. We model the fake news detection problem as a node classification task in the HIN, which requires us to learn the more comprehensive and discriminative representation of news article nodes.

The main challenges of the fake news detection problem in the HIN lie in the following three points:

- Hierarchy: Representation learning in heterogeneous networks will be a multi-level work, because the information of node contents and the information of the schema are contained at different levels.
- Heterogeneity: There exist various types of heterogeneous information related to news articles. Learning effective node representations in a HIN in a unified way is not an easy task.
- *Generalizability*: To ensure the applicability of the proposed model to different types of HINs, we need to propose a general learning model that can be extensible to various learning settings.

To handle these challenges aforementioned, we propose a novel Hierarchical Graph Attention Network (HGAT) to detect fake news. HGAT employs a hierarchical attention mechanism to learn the representation of news article nodes. Based on the learned node representation, fake news can be identified through the node classification task. In particular, for each news article node, we use the type-specific node-level attention mechanism to learn a set of weights for its neighbors with the same type. Using these sets of weights, we aggregate neighbor nodes of the same type into a schema node. The schema-level attention works to learn the attention weights of different schema nodes. Based on the two-level attention, HGAT can get the optimal combination of different types of neighbors in a hierarchical manner. The learned node representations capture the features from different heterogeneous information sources. HGAT can be optimized in an end-toend manner by backpropagation.

The contributions of our work are summarized as follows:

• To our best knowledge, this is the first attempt to detect the fake news in the heterogeneous information network without handcrafted features (e.g., meta-path).

- We propose the novel Hierarchical Graph Attention Network (HGAT) model, which takes different types of node contents and diverse categories of connections into consideration simultaneously. HGAT as a general model for representation learning has great potential to be applied to other applications in the HIN.
- We conduct extensive experiments on the real-world dataset to demonstrate the effectiveness of HGAT. The results show the superiority of HGAT comparing to the state-of-the-art models in detecting fake news.

2 Related Work

2.1 Fake News Detection

As an emerging topic, some research works have been proposed. Among them, the knowledge-based approach aims to assess the authenticity of news by comparing the knowledge extracted from the news contents with real knowledge [Ciampaglia et al., 2015]. Yet, the timeliness and integrity of the knowledge map remain an unresolved issue [Zhou and Zafarani., 2018]. Another typical way is based on writing style, such as discourse level by employing rhetorical structure theory [Rubin and Lukoianova., 2015], sentiment and readability [Perez-Rosas et al., 2017; Potthast et al., 2017]. Based on relationships among news articles, users (spreaders) and user posts, matrix factorization [Shu et al., 2019], tensor factorization [Gupta et al., 2018], hierarchical word encoder [Cui et al., 2019], and Recurrent Neural Networks (RNNs) [Ruchansky et al., 2017; Zhang et al., 2018] have been developed for fake news detection.

2.2 GNNs and Network Embedding

Graph Neural Networks (GNNs) for representation learning of graphs learn nodes' effective feature vectors through a recursive neighborhood aggregation scheme [Xu et al., 2019]. Kipf et al. [Kipf and Welling., 2017] propose Graph Convolutional Network (GCN). Graph Attention Network (GAT) [Velickovic et al., 2018] first imports the attention mechanism into graphs. However, the above graph neural networks are presented for the homogeneous graphs. Wang et al. [Wang et al., 2019] consider the attention mechanism in heterogeneous graph learning through the model HAN. However, meta-path as a handcrafted feature limits HAN, and HAN ignores the use of node contents carried by different types of nodes.

The learned embeddings from network embedding methods can be applied to the downstream tasks [Wang et al., 2019]. Some models have been proposed to deal with homogeneous networks including the random walk based methods [Hamilton et al., 2018], the matrix factorization based methods [Ou et al., 2016; Wang et al., 2017], the deep learning-based methods [Wang et al., 2016]. In order to handle the heterogeneity, metapath2vec [Dong et al., 2017] samples random walks under the guidance of meta-path on heterogeneous graphs. SHNE [Zhang et al., 2019] considers the heterogeneous networks as attributed graphs and jointly optimizes through heterogeneous SkipGram and semantic encoding.

3 Concept and Problem Definition

In this section, we first introduce some terminologies used in this paper and then formulate the studied problem.

3.1 Terminology Definition

DEFINITION 1 (News Articles): News articles refer to the news contents post on social media or public platforms. News articles can be represented as set $\mathcal{N} = \{n_1, n_2, \cdots, n_m\}$. For each news article $n_i \in \mathcal{N}$, it contains textual contents.

DEFINITION 2 (Subject): Subjects usually denote the central ideas of news articles, which are the main objectives of writing news articles. The set of subjects can be denoted as $S = \{s_1, s_2, \dots, s_k\}$. For each subject $s_i \in S$, it contains the textual description.

DEFINITION 3 (Creator): Creators denote users who write news articles. The set of creators can be represented as $C = \{c_1, c_2, \dots, c_n\}$. For each creator $c_i \in C$, it contains the profile information containing titles, political party membership, and geographical residential locations.

We can model News articles, Subjects and Creators into a heterogeneous network as three types of nodes and construct different types of links based on the connections among them. A formal definition of News Oriented Heterogeneous Information Networks can be proposed as follows:

DEFINITION 4 (News Oriented Heterogeneous Information Networks (News-HIN)): The news oriented heterogeneous information network (News-HIN) can be defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V} = \mathcal{C} \cup \mathcal{N} \cup \mathcal{S}$, and the link set $\mathcal{E} = \mathcal{E}_{c,n} \cup \mathcal{E}_{n,s}$ involves the "Write" links between creators and news articles, and the "Belongs to" links between news articles and subjects.

In order to better understand the News-HIN and utilize type information, it is necessary to define the schema-level description. The schema of News-HIN will be used in the model to learn the importance of nodes and links with different types.

DEFINITION 5 (News-HIN Schema): Formally, the schema of the given News-HIN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be represented as $S_{\mathcal{G}} = (\mathcal{V}_T, \mathcal{E}_T)$, where \mathcal{V}_T and \mathcal{E}_T denote the set of node types and link types in the network respectively. Here, $\mathcal{V}_T = \{\phi_n, \phi_c, \phi_s\}$ and $\mathcal{E}_T = \{Write, Belongs to\}$

We present the schema of News-HIN based on the PolitiFact dataset in Figure 1(b), where the exact node and link types can be found intuitively.

3.2 Problem Definition

Given a News-HIN $\mathcal{G}=(\mathcal{V},\mathcal{E})$, the fake news detection problem aims at learning a classification function $f:\mathcal{N}\longrightarrow\mathcal{Y}$ to classify news article nodes in the set \mathcal{N} into the correct class with the credibility label in \mathcal{Y} . Various kinds of heterogeneous information in the News-HIN \mathcal{G} should be effectively incorporated, including both the textual information and network structure information.

4 Proposed Method

Hierarchical Graph Attention Network (HGAT) follows a hierarchical attention structure including node-level attention and schema-level attention. The structure of HGAT is shown in Figure 2. The node-level attention is proposed to learn the weights of neighbors belong to the same type and aggregate them to get the type-specific neighbor representation. Then HGAT can learn the information of node types via schemalevel attention and achieve the optimal weighted combination for the final fake news detection task. We will discuss these components in this section.

4.1 Node-level attention

Node-level attention can learn the importance of neighbors belong to the same type respectively for each news article node $n_i \in \mathcal{N}$, and aggregate the representation of these meaningful neighbors to form an integrated representation which we define as a schema node.

The inputs of the node-level attention layer are the initial feature vectors of nodes. Because multiple types of nodes exist in the News-HIN, the initial feature vectors belong to feature spaces with different dimensions. In order to enable the attention mechanism to output comparable and meaningful weights between different types of nodes, we at first utilize a type-specific transformation matrix to project features with different dimensions into the same feature space. We take the news article node $n_i \in \mathcal{N}$ as an example. The transformation matrix for type ϕ_n is $\mathbf{M}^{\phi_n} \in \mathbb{R}^{F \times F^{\phi_n}}$, where F^{ϕ_n} is the dimension of the initial feature $h_{n_i} \in \mathbb{R}^{F^{\phi_n}}$ and F is the dimension of the feature space mapped to. The F is the same for all type-specific transformation matrices. The projection process can be shown as follows:

$$h'_{n_i} = \mathbf{M}^{\phi_n} \cdot h_{n_i}; h'_{c_i} = \mathbf{M}^{\phi_c} \cdot h_{c_i}; h'_{s_i} = \mathbf{M}^{\phi_s} \cdot h_{s_i}$$
 (1)

Through the type-specific projection operation, the feature space of nodes with different types can be unified, where the self-attention mechanism can work on to learn the weight among various kinds of nodes. Here, the node-level attention will learn the importance of same-type neighbor nodes respectively. In the face of detecting fake news, the target node is the news article node $n_i \in \mathcal{N}$ and the neighbors of it belong to $\mathcal{N} \cup \mathcal{S} \cup \mathcal{C}$. It should be noted that we also regard the target node itself as a neighbor node to cooperate the self-attention mechanism. We let $T \in \{\mathcal{N}, \mathcal{S}, \mathcal{C}\}$ and nodes in T have the same type ϕ_t , then for n_i 's neighbor nodes in T, the node-level attention can learn the importance $e_{ij}^{\phi_t}$ which means how important node $t_j \in T$ will be for n_i . The importance $e_{ij}^{\phi_t}$ can be formulated as follows:

$$e_{ij}^{\phi_t} = attention(h'_{n_i}, h'_{t_j}; \phi_t)$$
 (2)

Here, attention denotes the same deep neural network as [Velickovic et al., 2018] conducting the node attention. attention is shared for all neighbor nodes with the same type ϕ_t . The masked attention keeps the network structure information. Only node $t_j \in neighbor_{n_i}$ being neighbors of node n_i with the type ϕ_t will be calculated and recorded as $e_{ij}^{\phi_t}$.

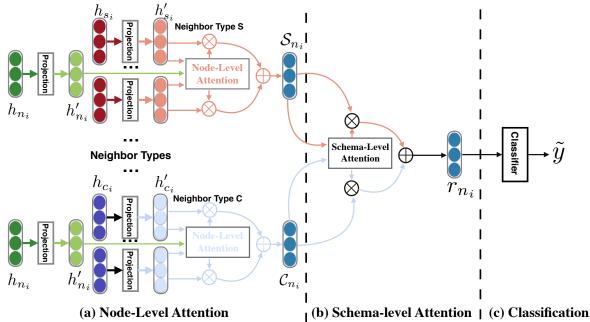


Figure 2: The overall framework of HGAT. (a) All types of nodes are projected into a unified feature space and the weights of node pairs can be learned via node-level attention. (b) Joint learning the weight of each type of schema nodes and fuse representations via schema-level attention. (c) Calculate the loss and end-to-end optimization.

Otherwise, the attention weight will be 0. We normalize them to get the weight coefficient $\alpha_{ij}^{\phi_t}$ via softmax function:

$$\alpha_{ij}^{\phi_t} = \operatorname{softmax}(e_{ij}^{\phi_t}) \tag{3}$$

Then, the schema node T_{n_i} can be aggregated as follows:

$$T_{n_i} = \sigma(\sum_{t_j \in neighbor_{n_i}} \alpha_{ij}^{\phi_t} \cdot h'_{t_j})$$
 (4)

Similar to Graph Attention Network (GAT) [Velickovic *et al.*, 2018], a multi-head attention mechanism can be used to stabilize the learning process of self-attention in node-level attention. In details, K independent node-level attentions execute the transformation of Equation (4), and then the features achieved by K heads will be concatenated, resulting in the output representation of the schema node:

$$T_{n_i} = \prod_{k=1}^{K} \sigma\left(\sum_{t_j \in neighbor_{n_i}} \alpha_{ij}^{\phi_t} \cdot h'_{t_j}\right)$$
 (5)

In the problem we face, every target node n_i has 3 schema nodes \mathcal{N}_{n_i} , \mathcal{C}_{n_i} , \mathcal{S}_{n_i} corresponding to 3 different types neighbors (include itself) based on the Definition 5.

4.2 Schema-level attention

Through the node-level attention, we aggregate the neighbors of news article nodes as several schema nodes. Essentially it is equivalent to fusing information from neighbor nodes of the same type into the representation of a schema node. What we still need to do now is to learn the representation of news article nodes from all schema nodes. Different schema

nodes contain type information, which require us to differentiate the importance of node types. Here we will use schemalevel attention to automatically learn the importance of different schema nodes, and finally use the learned coefficients for weighted fusion.

In order to obtain sufficient expressive power to calculate the attention weights between schema nodes as higher-level features, we will apply one learnable linear transformation to the features of schema nodes from node-level attention. The linear transformation is parametrized by a weight matrix $\mathbf{W} \in \mathbb{R}^{F' \times KF}$. K is the number of heads in node-level attention. The schema-leval attention mechanism schema is a single-layer feedforward neural network applying the activating function Sigmoid with the dimension 2F'. For the schema node T_{n_i} , the importance of it can be denoted as $w_i^{\phi_t}$:

$$w_i^{\phi_t} = schema(\mathbf{W}T_{n_i}, \mathbf{W}\mathcal{N}_{n_i}) \tag{6}$$

We normalize the imoportance of each schema nodes through a softmax function. Then coefficients of the final fusion are denoted as $\beta_i^{\phi_t}$, which can be calculated as follows:

$$\beta_i^{\phi_t} = \operatorname{softmax}(w_i^{\phi_t}) = \frac{\exp(w_i^{\phi_t})}{\sum_{\phi \in \mathcal{V}_T} \exp(w_i^{\phi})} \tag{7}$$

Based on the learned coefficients, we can fuse all schema nodes to get the final representation r_{n_i} of the target node n_i :

$$r_{n_i} = \sum_{\phi_t \in \mathcal{V}_T} \beta_i^{\phi_t} \cdot T_{n_i} \tag{8}$$

The set of learned final representation is denoted as \mathcal{R} . Figure 3 describes the two-level aggregating for reference.

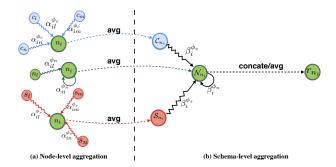


Figure 3: Explanation of aggregating process in both nodelevel and schema-level.

4.3 Loss Functions

Once achieving the final representation, we can use labeled news article nodes to train a classifier. In our experiments, a logistic regression layer is used to make predictions. We define the set of labeled news article nodes as \mathcal{N}_l . For the fake news detection tasks, our optimization objective function is set as a cross-entropy loss minimization and it can be optimized through the backpropogation.

In the binary-class fake news detection, the loss is:

$$Loss(\mathcal{R}, \mathcal{N}_l) = -\sum_{n_i \in \mathcal{N}_l} (y_{n_i} \log(p_{n_i}) + (1 - y_{n_i}) \log(1 - p_{n_i}))$$

$$(9)$$

Here, y is a binary indicator (0 or 1) indicating if the label is the correct classification for the news article node. p_{n_i} is the predicted probability of the representation of news article node n_i . The predicted probability will be output by a logistic regression layer in HGAT.

For the multi-class fake news detection, the cross-entropy based loss can be represented as:

$$Loss(\mathcal{R}, \mathcal{N}_l) = -\sum_{n_i \in \mathcal{N}_l} \sum_{j \in \mathcal{Y}} y_{n_i, j} \log(p_{n_i, j}) \qquad (10)$$

Where y is also a binary indicator (0 or 1) which indicates whether class label j is the correct classification for the news article node n_i . A multi-class logistic regression layer will be trained to output the predicted probability $p_{n_i,j}$ for each class. The overall algorithm of HGAT is described in Algorithm 1.

5 Experiments

To test the effectiveness of HGAT, extensive experiments are designed and conducted on the real-world fake news dataset. In this section, we first introduce the dataset. Then experimental settings and experimental results together with the detailed analysis are provided after that.

5.1 Dataset Description

Our dataset used in experiments is collected from the platform with fact-checking: *PolitiFact*, which is operated by Tampa Bay Times. Regarding news articles, *PolitiFact* provides the original contents, fact-checking results and comprehensive fact-checking reports on the website. The platform categorizes them into different subjects based on contents

Algorithm 1: HGAT

Input: The News-HIN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$;

```
The initial node feature h_i, i \in \mathcal{V}, \mathcal{V} = \mathcal{C} \cup \mathcal{N} \cup \mathcal{S}
                The News-HIN Schema S_{\mathcal{G}} = (\mathcal{V}_T, \mathcal{E}_T),
                \mathcal{V}_T = \{\phi_n, \phi_c, \phi_s\}
    Output: The learned representation r_{n_i}, n_i \in \mathcal{N};
                  The prediction labels vector \mathcal{L}
    begin
          for \phi_t \in \mathcal{V}_T do
2
3
                 for nodes t_i of the type \phi_t do
                      Feature space projection h'_{t_i} = \mathbf{M}^{\phi_t} \cdot h_{t_i};
4
          for n_i \in \mathcal{N} do
                 Find the neighbor nodes neighbor_{n_i};
6
                 for \phi_t \in \mathcal{V}_T do
7
                       for t_j \in neighbor_{n_i} do
8
                         Calculate the node-level coefficient \alpha_{ij}^{\phi_t};
                       Aggregate the schema node
10
                         T_{n_i} = \sigma(\sum_{t_i \in neighbor_{n_i}} \alpha_{ij}^{\phi_t} \cdot h'_{t_j})
          for n_i \in \mathcal{N} do
11
12
                for \phi_t \in \mathcal{V}_T do
13
                   Calculate the schema-level coefficient \beta_i^{\phi_t};
                 Aggregate to achieve learned representation
14
                   r_{n_i} = \sum_{\phi_t \in \mathcal{V}_T} \beta_i^{\phi_t} \cdot T_{n_i}
          Calculate Cross-Entropy and Back propagation;
15
          Update parameters and the prediction labels vector \mathcal{L};
16
          return r_{n_i}, n_i \in \mathcal{N}; \mathcal{L}
17
```

and topics. A brief description of each subject will be provided as well. The fact-checking results can indicate the credibility of corresponding news articles and take values from {True, Mostly True, Half True, Mostly False, False, Pants on Fire!}. In the *PolitiFact* dataset, 1322 news articles are marked as "Pants on Fire", while the number of news articles with "False" is 2601. Besides, 2539 "Mostly False" news articles and 2765 "Half True" news articles exist in the dataset. The number of "Mostly True" and "True" news is 2676 and 2149 respectively. If we group the labels {Pants on fire, False, Mostly False as fake news and group {True, Mostly True, Half True as real news, the quantity of fake news is 6465 corresponding to 7590 real news. We have established a heterogeneous information network based on the original dataset. The HIN includes three types of nodes: article, creator and subject and two types of links: Write (between article and creator) and Belongs to (between article and subject). The key statistical data describing the HIN can be found in Table 1.

Table 1: Properties of the Heterogeneous Networks

	property	PolitiFact Network
# node	article creator subject	14,055 3,634 152
# link	Write (creator-article) Belongs to (article-subject)	14,055 48,756

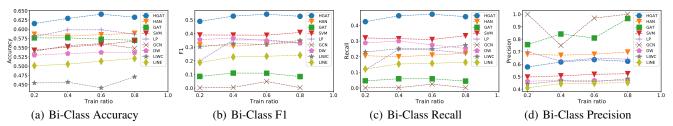


Figure 4: The results of bi-class news articles classification

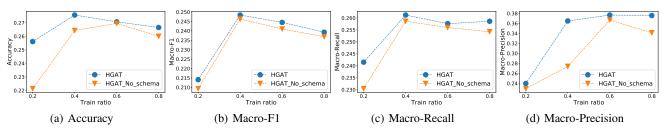


Figure 5: The comparison between HGAT and HGAT without schema-level attention

5.2 Experimental Settings

Experimental Setup

We are able to acquire the set of news article nodes which are the target nodes to conduct the classification. The set of news article nodes are divided into 10 folds. Among them, 8 folds will be used as the training set and 1 fold will be used as the validation set. The remaining 1 fold is left as the testing set. In order to conduct sufficient experiments with the setting of the different numbers of training data, we further make use of 2, 4, 6, 8 of 8 folds as the training set respectively. In this way, experiments will be conducted with training ratios $\theta \in \{20\%, 40\%, 60\%, 80\%\}$, and the testing ratio is fixed as 10%. The fact-checking results corresponding to news articles are used as the ground truth for model learning and evaluation. The node contents are encoded by TF-IDF to work as the initial feature vector of each type of nodes. We won't make use of comprehensive fact-checking reports in our experiments. We train models to work on both the multi-class classification task and the binary-class classification task. In the multi-class classification task, 6 kinds of different fack-checking results correspond to 6 classes. Meanwhile, in the binary-class classification, we group factchecking results {Pants on fire, False, Mostly False} as a Fake class and group {True, Mostly True, Half True} as a Real class. Because our target is to detect fake news, we treat Fake class as the positive class and Real class as the negative class. In the binary-class classification task, we evaluate the results with Accuracy, Precision, Recall, and F1. Meanwhile, when the model works on the multi-class classification task, the performance is evaluated by Accuracy, Macro Precision, Macro Recall, and Macro F1 respectively.

Comparison Methods

Graph neural network methods

• HAN [Yang et al., 2016]: HAN employs node-level at-

Table 2: The results of multi-class news articles classification

			Text-based		Network Embed		GNNs			
Train	Metric	I	SVM	LIWC	LP	DW	LINE	GAT	HAN	HGAT
20%	Accuracy	ī	0.1967	0.1432	0.2218	0.1932	0.1532	0.2110	0.2181	0.2561
	F1	Ī	0.1624	0.1225	0.1925	0.1562	0.0765	0.1054	0.1234	0.2141
- 1	Recall	Ī	0.1801	0.0965	0.2153	0.1718	0.1433	0.1975	0.1884	0.2415
- 1	Precision	I	0.1905	0.1409	0.2859	0.1742	0.0326	0.1687	0.2467	0.2397
	Accuracy	I	0.2042	0.1543	0.2278	0.1952	0.1567	0.2237	0.2240	0.2757
40%	F1	I	0.1775	0.1314	0.1944	0.1646	0.0798	0.1103	0.1441	0.2484
- 1	Recall	Ī	0.1892	0.0987	0.2183	0.1742	0.1505	0.1987	0.1853	0.2616
- 1	Precision	I	0.2047	0.1491	0.3037	0.1745	0.0401	0.1815	0.2572	0.3649
60%	Accuracy	Τ	0.2061	0.1513	0.2373	0.1969	0.1453	0.2214	0.2256	0.2707
	F1	T	0.1871	0.1321	0.2099	0.1647	0.0653	0.1162	0.1475	0.2445
- 1	Recall	I	0.1976	0.1002	0.2222	0.1764	0.1410	0.1954	0.1852	0.2576
- 1	Precision	Ī	0.2118	0.1561	0.2955	0.1966	0.0307	0.1870	0.2792	0.3767
-	Accuracy	I	0.2186	0.1567	0.2407	0.2013	0.1623	0.2212	0.2207	0.2665
80%	F1	Ī	0.1962	0.1305	0.2187	0.1669	0.0875	0.1037	0.1218	0.2393
- 1	Recall	Ī	0.2081	0.0954	0.2341	0.1830	0.1512	0.1975	0.1840	0.2586
- 1	Precision	I	0.2233	0.1553	0.3149	0.1896	0.0468	0.1819	0.2497	0.3757

tention and semantic-level attention to capture the information from all meta-paths. In our experiments, we utilize two meta-paths (article-creator-article, articlesubject-article) in HAN.

- **GAT** [Velickovic *et al.*, 2018]: GAT is an attention-based graph neural network for the node classification.
- **GCN** [Kipf and Welling., 2017]: GCN is a semi-supervised methods for the node classification.

Text-based methods

SVM: SVM is a basic supervised classification model.
 The feature vector used is extracted merely based on the news article contents with TF-IDF.

• LIWC [Pennebaker *et al.*, 2015]: LIWC is used to extract the lexicons falling into psycho-linguistic categories. We follow [Pennebaker *et al.*, 2015] to use LSTM to work o final news articles classification.

Network embedding methods

- Label Propagation (LP) [Zhu and Ghahramani, 2002]:
 LP is merely based on the network structure. The prediction scores of LP will be rounded and cast into labels.
- DeepWalk [B. Perozzi and Skiena, 2014]: A random walk based embedding method. Based on the embedding results, we then train a logistic regression model to perform the classification of news articles.
- LINE [Tang *et al.*, 2015]: LINE preserves the local and the global network structure simultaneously. We also learn a logistic regression model to conduct the classification based on the learned embeddings.

Among these baseline methods, **SVM**, **LIWC** use the text information only, while **DeepWalk**, **LINE** and **Label Propagation** learn from the network structure merely but ignore the heterogeneity. **GAT**, **GCN**, and **HAN** make use of textual contents and network structure simultaneously. We have also noticed some recently appeared methods for fake news detection [Cui *et al.*, 2019; Shu *et al.*, 2019], but did not compare them. The main consideration is the difference between the scenarios we face. In [Cui *et al.*, 2019; Shu *et al.*, 2019], they all utilize social context like user comments, but HGAT aims at detecting fake news in a relatively early stage. We won't utilize user comments about the news, because when many users have started to discuss one fake news, the bad influence of fake news has spread.

5.3 Reproducibility

For the proposed HGAT, we optimize the model with Adam. The dimension of node-level representations is set as 12 and the dimension of schema-level is set as (8 * K). Here, the attention head K is set as 1. For HAN, we set the dimension of node-level representations to 12 the same as HGAT, and the number of semantic-level hidden units is 8. For GAT, we set the embedding dimension as 12 and use just 1 attention head for a fair consideration. For GCN, the embedding dimension is set as 512. In the DeepWalk, we set the window size to 5, length of the random walk to 30, the number of walks per node to 10 and the embedding dimension to 128. We run the experiments on the Server with 3 GTX-1080 ti GPUs, and all codes are implemented in Python. The detailed parameters of all other comparison methods is provided in the project. Code is available at: https://github.com/YuxiangRen/Hierarchical-Graph-Attention-Network

5.4 Experimental Results with Analysis

Binary-class classification tasks

Based on results in Figure 4, our model HGAT achieves the best performance when focusing on Accuracy, F1 and Recall. However, when considering Precision, we can observe from Figure 4(d) that the performance of HGAT is lower than that of GCN and GAT. Through careful analysis, it can be found that GAT and GCN tend to judge most instances as 'Real' in

the face of fake news detection, so the higher precision is related to very low Recall. In this case, the higher precision is not practical, because a lot of false news can not be detected. By comparing the performance between HGAT and network embedding methods, we can conclude that textual information is quite important, and just based on network structure is insufficient. At the same time, through the comparison between HGAT and text classification methods, we can find that network structure is powerful to the fake news detection as well. At last, through the comparison among GNNs methods, we verify that the heterogeneity of networks should be dealt with in a more effective way. If we simply treat a heterogeneous network as a homogeneous network by ignoring the type, then the result would be very disappointing. Also as a method for heterogeneous graphs, HGAT also shows an advantage over HAN. More important. HGAT is a meta-pathfree model without the limitation of handcrafted features.

Multi-class classification tasks

Due to the uncertainty of the nature of emerging news, it is often difficult to judge news directly as absolutely true and false. Besides, it is also not conducive to subsequent operations (e.g., final verification). Carrying out finer granularity multi-classification tasks according to the credibility of news is very meaningful. The experimental results from 6-labels classification are shown in Table 2, where HGAT outperforms all comparison methods with an obvious advantage. From a more general point of view, this also shows that HGAT has a stronger learning ability in the heterogeneous network, and the learned representation is also more comprehensive and discriminative. The results demonstrate the great potential and scalability of HGAT in the face of other scenarios based in heterogeneous networks.

Performance of schema-level attention

In order to verify the effectiveness of schema-level attention, we replace the schema-level attention of HGAT with fixed and equal weights for schema nodes. In experiments, all three schema nodes are assigned with the weight 1/3, and we denote this comparison model as HGAT_No_schema. In Figure 5, we present the comparison results between HGAT and HGAT_No_schema. The results come from the experiments in a multi-class classification setting with different train ratios. It's obvious that HGAT achieves better performance than HGAT_No_schema according to various metrics. This comparison illustrates that the importance of schema nodes worth distinguishing, and HGAT can differentiate the importance through attention weights effectively. In contrast, the simple average operation to aggregate schema nodes harms the performance, which is equivalent to droping the type information of schema nodes in essence.

6 Conclusion

In this paper, we study the HIN-based fake news detection problem and propose a novel graph neural network HGAT to solve it. Based on the News-HIN, textual information regarding news articles, area background and creator profiles along with the network structure information can be captured. HGAT employs a hierarchical attention mechanism considering both node-level and schema-level attention to learn the comprehensive representations of news article nodes. These effective and discriminative representations can be used to detect fake news. HGAT is also a general graph representation learning model that does not require any handcrafted features like meta-path or other prior knowledge, so it is highly extensible for heterogeneous network-based problems other than fake news detection. Extensive experiments are conducted on a real-world News-HIN, i.e., PolitiFact. The experiment results demonstrate that HGAT has outstanding performance compared with the state-of-the-art methods.

References

- [Akoglu et al., 2013] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews by network effects. In ICWSM, 2013.
- [Allcott and Gentzkow, 2017] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 2017.
- [B. Perozzi and Skiena, 2014] Rami Al-Rfou
 B. Perozzi and Steven Skiena. Deepwalk:
 Online learning of social representations. In KDD, 2014.
- [Ciampaglia et al., 2015] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 2015.
- [Cui et al., 2019] Limeng Cui, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. defend: A system for explainable fake news detection. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pages 2961–2964. ACM, 2019.
- [Dong *et al.*, 2017] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. Meta-path guided embedding for similarity search in large-scale heterogeneous information networks. In *KDD*, 2017.
- [Gupta et al., 2018] Shashank Gupta, Raghuveer Thirukovalluru, Manjira Sinha, and Sandya Mannarswamy. Cimtdetect: A community infused matrix-tensor coupled factor- ization based method for fake news detection. In arXiv preprint arXiv:1809.05252, 2018.
- [Hamilton *et al.*, 2018] Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs. In *NIPS*, 2018.
- [Jin *et al.*, 2017] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, and J. Luo. Detection and analysis of 2016 us presidential election related rumors on twitter. In *SBP-BRiMS*, 2017.

- [Kipf and Welling., 2017] Thomas N. Kipf and Max Welling. Semi-supervised classi cation with graph convolutional networks. In *ICLR*, 2017.
- [Ou *et al.*, 2016] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. Asymmetric transitivity preserving graph embedding. In *KDD*, 2016.
- [Pennebaker et al., 2015] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc. *Technical Report*, 2015.
- [Perez-Rosas et al., 2017] Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In arXiv preprint arXiv:1708.07104, 2017.
- [Potthast *et al.*, 2017] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendor, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *arXiv preprint arXiv:1702.05638*, 2017.
- [Rubin and Lukoianova., 2015] Victoria L Rubin and Tatiana Lukoianova. Truth and deception at the rhetorical structure level. *Journal of the* Association for Information Science and Technology, 2015.
- [Ruchansky et al., 2017] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In CIKM, 2017.
- [Shi *et al.*, 2017] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. Yu. A survey of heterogeneous information network analysis. *TKDE*, 2017.
- [Shu *et al.*, 2019] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *WSDM*, 2019.
- [Sun and Han, 2012] Y. Sun and J. Han. Mining heterogeneous information networks: a structural analysis approach. *KDD Explorations*, 2012.
- [Tang et al., 2015] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In WWW, 2015.
- [Velickovic *et al.*, 2018] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. In *ICLR*, 2018.
- [Wang et al., 2016] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In KDD, 2016.
- [Wang et al., 2017] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In AAAI, 2017.

- [Wang et al., 2019] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, P. Yu, and Yanfang Ye. Heterogeneous graph attention network. In WWW, 2019.
- [Xie *et al.*, 2012a] S. Xie, G. Wang, S. Lin, and P. Yu. Review spam detection via temporal pattern discovery. In *KDD*, 2012.
- [Xie et al., 2012b] S. Xie, G. Wang, S. Lin, and P. Yu. Review spam detection via time series pattern discovery. In WWW, 2012.
- [Xu et al., 2019] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.
- [Yang et al., 2016] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In NAACL, 2016.
- [Zhang et al., 2018] Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B Gouza. Fake news detection with deep diffusive network model. In arXiv preprint arXiv:1805.08751, 2018.
- [Zhang *et al.*, 2019] Chuxu Zhang, Ananthram Swami, and Nitesh V. Chawla. Shne: Representation learning for semantic-associated heterogeneous networks. In *WSDM*, 2019.
- [Zhou and Zafarani., 2018] Xinyi Zhou and Reza Zafarani. Fakenews: A survey of research, detection methods, and opportunities. In *arXiv* preprint arXiv:1812.00315, 2018.
- [Zhu and Ghahramani, 2002] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, 2002.