TaHiD: Tackling Data Hiding in Fake News Detection with News Propagation Networks

Anonymous ACL submission

Abstract

Fake news with detrimental societal effects has attracted extensive attention and research. Despite early success, the state-of-the-art methods fall short of considering the propagation of news. News propagates at different times through different mediums, including users, comments, and sources, which form the news propagation network. Moreover, the serious problem of data hiding arises, which means that fake news publishers disguise fake news as real to confuse users by deleting comments that refute the rumor or deleting the news itself when it has been spread widely. Existing methods do not consider the propagation of news and fail to identify what matters in the process, which leads to fake news hiding in the propagation network and escaping from detection. Inspired by the propagation of news, we propose a novel fake news detection framework named TaHiD, which models the propagation as a heterogeneous dynamic graph and contains the propagation attention module to measure the influence of different propagation. Experiments demonstrate that TaHiD extracts useful information from the news propagation network and outperforms state-of-the-art methods on several benchmark datasets for fake news detection. Additional studies also show that TaHiD is capable of identifying fake news in the case of data hiding.

1 Introduction

Nowadays, social media has been widely used. As Statista[1] reported, there were nearly 300 million social media users in the United States in 2021. Due to the widespread use of social media, more and more people use it to obtain and disseminate news. About half of U.S. adults (53%) say they get news from social media "often" or "sometimes", and this use is spread out across several different sites, according to a Pew Research Center survey[2]. However, at the same time, social media has made it easier to spread misinformation and disinformation, especially **fake news**. A majority of the Americans who are getting news on social media continue to question its accuracy. About six in ten (59%) of those who at least rarely get news on social media say they expect that news to be largely inaccurate[2]. Fake news contains intentional false information and can disrupt social order. For example, Cui and Lee (2020) mentioned that 77 cell phone towers have been set on fire due to the conspiracy claiming that 5G mobile networks can spread COVID-19. Since widespread fake news has harmful social efforts, there is an urgent need for developing fake news detection methods.

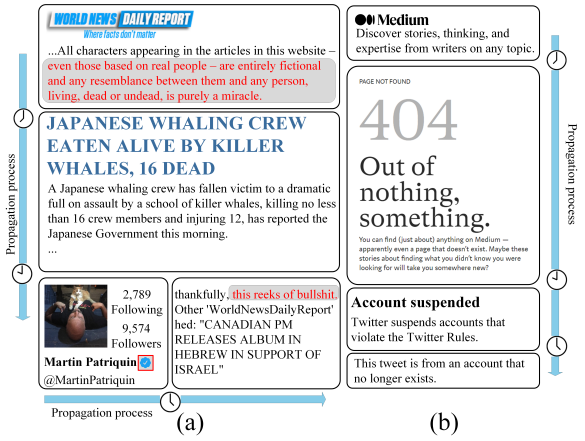


Figure 1: A diagram of news propagation. The blue arrows represent the news propagation process. (a) Besides news content itself, the entities in the process of propagation are important. Here is an example of fake news and its propagation mediums, where the highlight parts help classify this news as fake. (b) The figure shows an example of data hiding. During the propagation of news, the information even the news itself would be deleted by its publisher to be hidden.

Previous fake news detection works mainly

[1]https://www.statista.com/statistics/278341/number-of-social-network-users-in-selected-countries/

[2]https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/

adopted traditional feature engineering. Pennebaker et al. (2015) extracted features from the psychology and deception perspective for text classification. Castillo et al. (2011) adopted features derived from text like the number of positive sentiment words and the number of URLs in the news to detect fake news. Since feature engineering has serious subjective errors, researchers began to leverage neural network based fake news detection frameworks. Shu et al. (2019a) provided a way to exploit both news contents and user comments by a hierarchical attention neural network jointly to detect fake news. Wang et al. (2018) proposed a framework that contains VGG19 (Simonyan and Zisserman, 2014) to extract multi-modal features of news. As the research progressed, researchers realized that the social network composed of news is helpful to improve fake news detection performance. Shu et al. (2020b) built a news propagation network and extracted macro-level and micro-level network features to identify fake news. Nguyen et al. (2020) adopted graph neural networks to learn news representations and applied the learned representations to detect fake news.

Despite early successes, previous methods have failed to identify fake news in the news propagation networks as Figure 1 illustrates. Most of the earlier work focused only on the news itself, ignoring the process of news propagation. The entities in the news propagation networks contain a lot of information. For example, Shu et al. (2019b) point out that users who share news can help detect fake news. The more serious problem is data hiding, which means fake news publishers disguise fake news as real to confuse people. They hide fake news by deleting debunking comments or news itself. Fake news may hide in the propagation network and escape from detection. In light of this, there is an urgent need for a method that considers the propagation process of news and tackles the data hiding problem.

Inspired by the propagation process, we propose TaHiD (**Ta**ckling Data **Hi**ding in Fake News **D**etection with News Propagation network), a fake news detection framework. TaHiD considers the news propagating process as a heterogeneous dynamic graph and encodes critical entities of the propagation process, including news, sources, comments, users, and temporal information. TaHiD contains a propagation attention module to identify important propagation in the process, which ensures that TaHiD is still effective in the absence of propagation network information. Our main contributions are summarized as follows:

- We propose an end-to-end fake news detection framework named TaHiD. TaHiD considers news propagation networks as a heterogeneous dynamic graph, from which extracts information to identify fake news.
- TaHiD contains the propagation attention module, which measures the contributions of each propagation, to address the data hiding problem.
- We conduct extensive experiments on three real-world datasets to evaluate TaHiD and competitive baselines, which shows the excellent ability of TaHiD to identify fake news.

2 Related Work

The widespread dissemination of fake news on social media has brought serious harm to the politic, economy, and other fields. Researchers adopt various method on fake new detection (Zellers et al., 2019; Fung et al., 2021; Dementieva and Panchenko, 2021). Previous works on fake news detection mainly focus on text features. Castillo et al. (2011) use features from users' posting and retweeting behavior, tweet content and URLs. Popat et al. (2016) propose a classifier that uses factors to assess the credibility of the claim from different sources. Deep Neural Networks are also adopted recently and significantly outperform traditional methods. Karimi and Tang (2019) utilize automatic document structure learning and learn structurally rich representations for news documents. Ma et al. (2018) integrate both structure and content semantics based on tree-structured recursive neural networks for detecting rumors. Tan et al. (2020) introduce the task of detecting news generated by machines which includes images. Pelrine et al. (2021) find that traditional NLP baselines are competitive with and can outperform novel transformer-based methods. However, due to the arbitrary size and topology of the social graph, performing CNNs on graphs is not a viable solution.

GNNs are neural networks that can be directly applied to graphs, which provide an easy way to perform node-level, edge-level, and graph-level prediction tasks. Many graph-based efforts were made on the task of fake news detection (Monti et al., 2019; Gangireddy et al., 2020; Zhong et al.,

2020; Benamira et al., 2019; Wang et al., 2020). Lu and Li (2020) develop GCAN which learns the representations of user interactions, retweet propagation, and their correlation with source short text. Rossi et al. (2020) design TGN, which is a generic, efficient framework for dynamic graphs represented as sequences of timed events. Feng et al. (2021) propose RGT, which aims to leverage the heterogeneity in Twitter networks and counter misinformation and bots. Hu et al. (2021) utilize external knowledge through entities for fake news detection. Han et al. (2020) use the propagation pattern of news on social media and focus on a propagation-based approach for fake news detection. Mehta and Goldwasser (2021) contribute a novel benchmark for fake news detection at the knowledge element level, as well as a solution for this task which incorporates cross-media consistency checking to detect the fine-grained knowledge elements and make news articles misinformative. Mehta et al. (2022) formulate fake news detection as a reasoning problem and propose an inference-based graph representation learning method. Works on fake news detection often simplified social graphs while real-world scenarios are dynamic, heterogeneous, and more complicated.

3 Methodology

Figure 2 shows the overview of TaHiD, which consists of four components: (i) news propagation network; (ii) heterogeneous dynamic graph encoding module; (iii) propagation attention module; (iv) prediction module.

3.1 News Propagation Network

Let us first define the propagation network $G = G(A, S, C, U, R)$ with its entities and relations as Figure 2(i) illustrates. We denote $A = \{a_1, a_2, \cdots, a_{N_A}\}$ as the news entities, where a_i is i-th news entity and N_A is the number of news. Similarly, $S = \{s_1, s_2, \cdots, s_{N_s}\}$, $C = \{c_1, c_2, \cdots, c_{N_c}\}$ and $U = \{u_1, u_2, \cdots, a_{N_U}\}$ represent source, comment and user respectively. $R = \{e_1, e_2, \cdots, e_M\}$ is the list of relations in G. $e_i = (e_i^s, e_i^t, e_i^{type})$ is considered as the relation between the source entity e_i^s and the target entity e_i^t, where $e_i^s, e_i^t \in A \cup S \cup C \cup U$. $e_i^{type} \in R_{edge}$ is the type of this edge and R_{edge} is the set of types. More detailed information about news propagation network can be found in Appendix. We can now formally define the fake news detection using news propagation network.

Definition. *Given a propagation network $G = G(A, S, C, U, R)$ constructed from news A, news sources S, related comments C, users U, and edges R, our task is constructing a classification function $f : f(G(A, S, C, U, R)) \rightarrow \hat{y}$, where $\hat{y}$ is the predicted label of each news $a_i \in A$, such that $\hat{y}$ approximates ground truth y to maximize prediction accuracy.*

3.2 Heterogeneous Dynamic Graph Encoding

For simplicity, we omit the subscript standing for each entity in the following detail.

News encoding TaHiD encodes news title and content, then concatenates them to form an overall feature vector for news, *i.e.*,

$$r_a = r_a^{title} \| r_a^{content} \in \mathbb{R}^D, \quad (1)$$

where D is a hyperparameter denoting the news embedding dimension.

The most important entity in news is title, which leads the outline and attracts the audience. TaHiD derive the representation vector of news title as

$$r_a^{title} = \phi(W_A^T \cdot (\frac{1}{Q_T} \sum_{j=1}^{Q_T} \overline{a}_j^{title}) + b_A^T), \quad (2)$$

where $\phi(\cdot)$ is the activate function, $W_A^T \in \mathbb{R}^{D_s \times D/2}$, $b_A^T \in \mathbb{R}^{D/2}$ are learnable parameters, D_s is word embedding dimension, and $\overline{a}_j^{title} \in \mathbb{R}^{D_s}$ denotes the word embedding derived from the pre-trained language model RoBERTa (Liu et al., 2019) with considerable language modeling capabilities, calculated by

$$\{\overline{a}_j^{title}\}_{j=1}^{Q_T} = \text{RoBERTa}\{a_j^{title}\}_{j=1}^{Q_T}, \quad (3)$$

where a_j^{title} is j-th word in title tokenized by NLTK (Bird et al., 2009) while Q_T is the word count.

As fake news is created to spread inaccurate and harmful information, their content often have language style different from normal ones. To comprehensively represent the content information of news, TaHiD derives a single representation vector from news content as

$$r_a^{content} = \phi(W_A^C \cdot (\frac{1}{Q_S} \sum_{i=1}^{Q_S} \overline{a}_i^{sent}) + b_A^C), \quad (4)$$

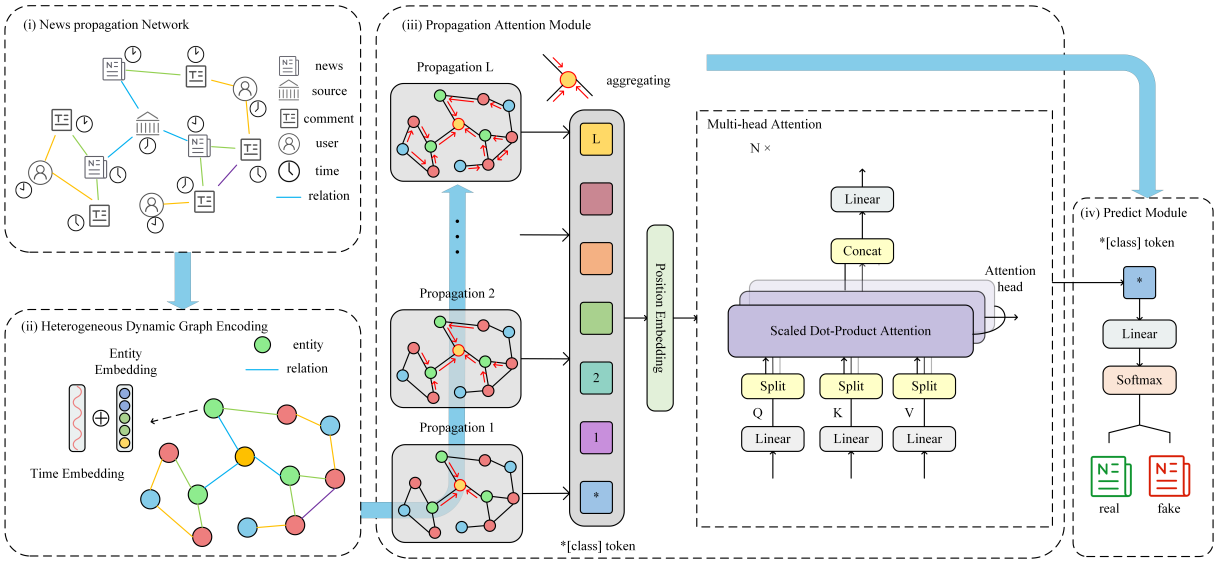


Figure 2: Overview of our proposed framework TaHiD.

where $W_A^C \in \mathbb{R}^{D/2 \times D_s}$, $b_A^C \in \mathbb{R}^{D/2}$ are learnable parameters, and $\overline{a}_i^{sent}$ is the sentence representation averaged from word representation, calculated by

$$\{\overline{a}_i^{sent}\}_{i=1}^{Q_S} = \frac{1}{Q_i}\sum_{j=1}^{Q_i} \overline{a}_{i,j}^{word}, \overline{a}_i^{sent} \in \mathbb{R}^{D_s}, \quad (5)$$

where $\overline{a}_{i,j}^{word} \in \mathbb{R}^{D_s}$ denotes the word representation transformed with RoBERTa, *i.e.*,

$$\{\overline{a}_{i,j}^{word}\}_{j=1}^{Q_i} = \text{RoBERTa}(\{a_{i,j}^{word}\}_{j=1}^{Q_i}), \quad (6)$$

where $\{a_{i,j}^{word}\}_{j=1}^{Q_i}$ is tokenized by NLTK from i-th sentence $\{a_i^{sent}\}$, Q_i is the count of words; $\{a_i^{sent}\}_{i=1}^{Q_S}$ is tokenized from news content, and Q_S is the count of sentences.

Source encoding TaHiD encodes news source using its description in the homepage, *i.e.*,

$$r_s = r_s^{description} \in \mathbb{R}^D. \quad (7)$$

News sources are a critical component in news propagation. News published on well-known news sources will attract the attention and forward other relevant news media or personal media, and then achieve the goal of propagation. Similarly to news content representation, TaHiD constructs the source description feature vector $r_s^{description} \in \mathbb{R}^D$ using RoBERTa.

User encoding Users play an important role in spreading fake news, and therefore it is a good idea to use user's feature to detecting fake news. TaHiD encodes user to r_u as

$$r_u = r_u^{description} \| r_u^{property} \in \mathbb{R}^D. \quad (8)$$

The rich semantic information in user description is helpful to identify a user and demonstrate the influence in spreading news. In TaHiD, RoBERTa is also adopted to encode a user's description $r_u^{description} \in \mathbb{R}^{D/2}$.

Properties refer to the statistics of users such as follower count or whether the user is verified, which are widely exploited in different tasks on social media. TaHiD adopts the properties which can be obtained from Twitter API[3] directly. TaHiD conducts z-score normalization for numerical properties (e.g. favourites count) while $0-1$ encoding for true-or-false categorical properties (e.g. verified), then feeds these raw properties into MLPs to construct $r_u^{property} \in \mathbb{R}^{D/2}$.

Comment encoding Users express their emotions or opinions towards news through posting comments on social media. These comments are likely related to the original news, which is helpful for fake news detection. TaHiD encodes comments using its textual content and properties, *i.e.*,

$$r_t = r_t^{content} \| r_t^{property} \in \mathbb{R}^D. \quad (9)$$

TaHiD adopts method similar to constructing news content feature vector and user properties feature vector to get $r_t^{content} \in \mathbb{R}^{D/2}$ and $r_t^{property} \in \mathbb{R}^{D/2}$, individually.

Time encoding TaHiD considers the news propagation network as a dynamic network, where each entity has a creation timestamp. Temporal information like the publication time of news or the post

[3]https://developer.twitter.com/en/docs/twitter-api

time of comment can help to detect fake news. To capture the potential temporal relation among entities, TaHiD adopts the temporal embedding. We believe that the temporal information of the different types of edges has different effects. TaHiD adds time encoding relative to the entity t to its own representation r, *i.e.*,

$$\tilde{r} = r + \sum_{e^{type} \in R_{edge}} r_{time}^{t,e^{type}}, \tag{10}$$

where $\tilde{r}$ denotes the initial representation of this entity while $r_{time}^{t,e^{type}}$ denotes the temporal embedding of entity t for each edge type e^{type}, which is calculated by

$$\begin{aligned} r_{time}^{t,e^{type}} &= \text{mean}(\{r_{time}^{s,t,e^{type}}\}_{(s,t,e^{type}) \in E}), \\ r_{time}^{s,t,e^{type}} &= \text{embed}(|s^{time} - t^{time}|, e^{type}), \end{aligned}$$

where $r_{time}^{s,t,e^{type}}$ means the influences of a edge (s, t, e^{type}) to entity t, $|\cdot|$ is absolute value operation, $\text{mean}(\cdot)$ is average value operation, and $\text{embed}(\cdot,\cdot)$ is the time embedding operation. We construct the time embedding operation inspired by Hu et al. (2020) as follows

$$\begin{aligned} &\text{Base}(\Delta t, 2i) = sin(\Delta t / 10000^{\frac{2i}{D}}), \\ &\text{Base}(\Delta t, 2i+1) = cos(\Delta t / 10000^{\frac{2i+1}{D}}), \\ &\text{embed}(\Delta t, e^{type}) = W_T^{type} * \text{Base}(\Delta t) + b_T^{type}, \end{aligned}$$

where W_T^{type} and b_T^{type} are learnable parameters, and Δt means the time difference.

3.3 Propagation Attention Module

TaHiD considers representation of each entity from Equation (10) as the first propagation vectors $X^{(1)} \in \mathbb{R}^{N \times D}$, where $N = N_A + N_S + N_U + N_T$. TaHiD leverages graph neural network as propagation function to obtain $(l+1)$-th propagation vectors $X^{(l+1)}$ from l-th propagation vectors $X^{(l)}$ and adjacency matrix A from edges set R. TaHiD adopts (i) **GCN** (Kipf and Welling, 2016); (ii) **GAT** (Velickovic et al., 2017); (iii) **HGT** (Hu et al., 2020); (iv) **R-GCN** (Schlichtkrull et al., 2018) as different propagation function.

Considering that the propagation vectors from different layers contribute differently, TaHiD utilizes a propagation attention module to aggregate information from different propagations. For an entity in the propagation networks, assume that its different propagation vectors are $\{x^{(1)}, x^{(2)}, \cdots, x^{(l)}\}$. The sequence of tokens input to the following transformer encoder is defined as Z, *i.e.*,

$$Z = [x^{(0)}, x^{(1)}, \cdots, x^{(l)}] + emb_p \in \mathbb{R}^{(l+1) \times D},$$

where $x^{(0)}$ is a learnable embedding to the sequence of propagation representation similar to BERT (Devlin et al., 2018)'s [class] token and emb_p is a learnable position embedding. TaHiD takes the first token, namely [class] token, from $\tilde{Z}$ as the final representation z of this entity, denoted by

$$z = \tilde{Z}[0], \tag{11}$$

where $\tilde{Z}$ are the representations of tokens, which are obtained by concatenating the output of each head and putting them into MLPs, *i.e.*,

$$\tilde{Z} = \text{MLPs}([head_1, \cdots, head_p]), \tag{12}$$

where p is the count of heads. The attention operators (Vaswani et al., 2017) of the i-th head $head_i$ is defined as

$$head_i = \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d_k}}) V_i, \tag{13}$$

where $Q_i = ZW_{Q,i}$, $K_i = ZW_{K,i}$, $V_i = ZW_{V,i}$ and $W_{Q,i}, W_{K,i}, W_{V,i} \in \mathbb{R}^{D \times d_k}$ are learnable parameters. $d_k = D/p$ is the head dimension.

3.4 Prediction Module

For each news, the goal is to minimize the loss function, *i.e.*,

$$L(\theta) = -ylog(\hat{y}) - (1-y)log(1-\hat{y}), \tag{14}$$

where θ denotes the learnable parameters of TaHiD, y is the ground truth of this news, and $\hat{y}$ is the predicted probability of fake news. The probability $\hat{y}$ is obtained by a softmax layer based on news' representation from Equation (11), *i.e.*,

$$\hat{y} = \text{Softmax}(W * z + b), \tag{15}$$

where $W \in \mathbb{R}^{D \times 2}$ and $b \in \mathbb{R}^2$ are learnable parameters.

4 Experiments

4.1 Experiment Settings

Dataset We make use of three real-world datasets to evaluate TaHiD and baselines.

- **Politifact** (Shu et al., 2020a) The dataset is collected from the *PolitiFact* platform.

Table 1: Performance comparison for fake news detection methods. "/" denotes insufficient dataset information to support the baseline[4]. We perform five experiments for each baseline and report the mean and standard deviation.

Methods		Politifact			Gossipcop			CoAID		
		F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC
HAN		72.5 (±1.0)	78.5 (±1.1)	83.9 (±2.2)	65.6 (±0.8)	86.7 (±0.3)	86.4 (±0.3)	88.3 (±1.6)	96.2 (±0.5)	98.5 (±0.7)
EANN		69.8 (±3.3)	78.3 (±2.2)	84.9 (±0.8)	64.9 (±1.8)	86.8 (±0.5)	86.9 (±0.7)	86.9 (±0.7)	95.8 (±0.2)	98.9 (±0.2)
dEFEND		79.7 (±3.2)	84.5 (±2.4)	91.9 (±2.6)	77.8 (±2.8)	90.3 (±1.0)	94.9 (±1.0)	89.1 (±2.7)	96.5 (±0.8)	99.3 (±0.2)
PNUP		82.7 (±0.7)	87.0 (±0.6)	86.0 (±0.5)	86.8 (±0.3)	95.5 (±0.1)	92.8 (±0.1)	/	/	/
RoBERTa		87.4 (±0.9)	90.3 (±0.8)	96.1 (±0.2)	64.2 (±0.5)	85.7 (±0.4)	85.8 (±0.2)	88.4 (±0.7)	96.1 (±0.3)	99.3 (±0.0)
Cross-Domain		83.9 (±2.1)	88.6 (±1.5)	94.6 (±0.6)	79.5 (±1.4)	90.7 (±0.7)	94.9 (±1.0)	82.5 (±2.5)	94.3 (±0.9)	97.4 (±0.3)
TaHiD	GCN	88.8 (±1.4)	91.4 (±1.2)	97.6 (±0.5)	94.9 (±0.6)	97.8 (±0.2)	99.6 (±0.0)	94.2 (±0.5)	98.1 (±0.2)	**99.5** (±0.1)
	HGT	**91.2** (±0.8)	**93.5** (±0.7)	97.1 (±1.1)	72.6 (±0.7)	88.6 (±0.3)	91.2 (±0.2)	88.8 (±1.0)	96.4 (±0.4)	98.7 (±0.0)
	GAT	89.2 (±1.3)	92.1 (±1.1)	**97.8** (±0.6)	**95.4** (±0.6)	**98.0** (±0.3)	99.6 (±0.2)	94.7 (±0.4)	98.2 (±0.2)	99.4 (±0.2)
	R-GCN	89.4 (±0.7)	92.2 (±0.9)	96.8 (±1.0)	95.2 (±0.5)	97.9 (±0.2)	**99.7** (±0.0)	**94.7** (±0.2)	**98.2** (±0.1)	99.4 (±0.2)

- **Gossipcop** (Shu et al., 2020a) This dataset is collected from the *GossipCop* platform.

- **CoAID** (Cui and Lee, 2020) This dataset includes diverse COVID-19 healthcare misinformation, including fake news on websites and social platforms.

These three datasets contain news content with labels. For each news, we collect related social network information including users, comments on Twitter, and source information. We randomly conduct a 7:2:1 partition for three datasets as training, validation, and test set.

Baseline methods We compare TaHiD with the following methods as baselines:

- **HAN** (Yang et al., 2016) adopts a hierarchical attention neural network framework on news contents for fake news detection.

- **EANN** (Wang et al., 2018) is a general framework for fake news detection that contains an integrated multi-modal feature extractor.

- **dEFEND** (Shu et al., 2019a) is an explainable fake news detection framework that exploits both news content and user comments jointly.

- **UPFND** (Shu et al., 2019b) characterizes and understands user profile features to classify fake news.

- **RoBERTa** (Liu et al., 2019) uses the pre-trained weight to derive news content features and adopts a fully-connect layer to classify fake news.

- **Cross-domain** (Silva et al., 2021) is a multi-modal fake news detection technique that learns domain-specific and cross-domain information of news records.

Evaluation metrics We adopt F1-score, Accuracy, and AUC as evaluation metrics of different fake news detection methods.

Implement We implement TaHiD framework with pytorch (Paszke et al., 2019), pytorch geometric (Fey and Lenssen, 2019), and the transformers library (Wolf et al., 2020). We submit our code and detailed information as part of the supplementary material. More detailed information about the implementation can be found in the appendix.

4.2 Fake News Detection Performance

Table 1 reports fake news detection performance of different methods on three datasets, which demonstrates that:

- TaHiD based methods achieve the best performance compared with other baselines over all of the datasets. TaHiD achieves about 4.35%, 8.66%, and 6.29% relative performance improvement on the F1-score, respectively.

- TaHiD with HGT achieves the best performance on Politifact while TaHiD with R-GCN achieves the best on CoAID. They achieve 2.70% and 0.53% improvement compared to TaHiD with GCN, which illustrates that heterogeneous information can help identify fake news.

- Methods only consider the news content, such as HAN, only get an F1-score of 65.6% on Gossipcop. It is shown that news content is not enough for fake news detection.

- Cross-Domain achieves better performance on Politifact with an F1-score of 83.9% and Gossipcop with an F1-score of 79.5%, which suggests

[4]PNUP adopts user information to identify fake news while CoAID contains little user information.

Table 2: The performance of TaHiD time encoding ablation study on dataset Politifact. We train without time encoding. The "Type" column denotes the propagation function of TaHiD.

Type	F1			Acc		
	Prev.	w/o	Diff.	Prev.	w/o	Diff.
GCN	88.8	88.1	-0.7	91.4	90.7	-0.7
HGT	91.0	90.0	-1.0	93.3	92.5	-0.8
GAT	89.2	88.1	-1.1	92.1	90.7	-1.4
R-GCN	89.4	88.3	-1.1	92.1	91.6	-0.5

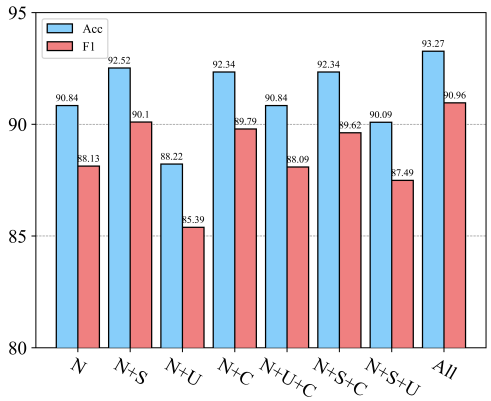


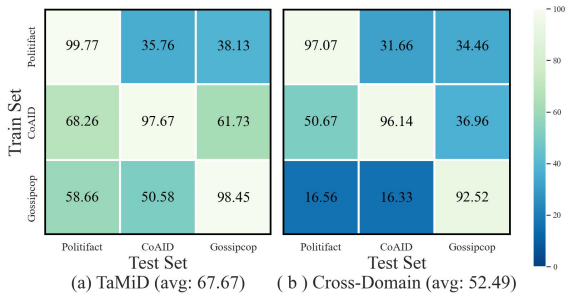
Figure 3: The performance of TaHiD entity ablation study on dataset Politifact. The "N", "S", "U", and "C" denote retaining news, source, user, and comment entity to train TaHiD while "All" denote using the whole propagation network.

that graph information can help improve performance. TaHiD considers the graph propagation, which leads to the best performance.

- PNUP achieves the second-best performance on Gossipcop with an F1-score of 86.8% but worse on Politifact with an F1-score of 82.7%, which shows that user information is helpful on the specific dataset. TaHiD considers the different contributions of every propagation, which leads to the best performance on three datasets.

4.3 TaHiD Propagation Study

Entity study TaHiD adopts various critical entities, namely user, comment, and source information to detect fake news. To figure out whether the idea of using such information has led to better performance, we conduct an ablation study that removes one kind of entity in the news propagation network at a time. Results in Figure 3 show that removing any entity in the news propagation network from TaHiD would lead to a considerable loss in performance. F1-score of TaHiD trained with news and comment drops from 90.96% to 89.79%, a relative drop of 1.29%. TaHiD trained with news and user drops the most, with a relative drop of 6.12% on the F1-score. It indicates that comment information is helpful for fake news detection while user information is not quite useful on Politifact. TaHiD extracts the most critical information from the propagation network to improve the performance of fake news detection.



(a) TaMiD (avg: 67.67) (b) Cross-Domain (avg: 52.49)

Figure 4: The results of training TaHiD on one dataset and testing on the other datasets.

Time embedding study TaHiD considers the news propagation network as a heterogeneous dynamic graph and adopts temporal encoding to get the temporal information of each entity and relation. To demonstrate the effect of TaHiD on extracting temporal information, we conduct an ablation study that removes the temporal encoding. From Table 2, we could find that removing time encoding leads to a drop in performance. The performance of TaHiD with GAT drops the most with 1.23% on the F1-score, while TaHiD with GCN drops the least with 0.79%. It proves that TaHiD improves the performance by encoding time information.

Transfer study To further prove TaHiD's ability to extract propagation network information, we train TaHiD and a competitive baseline, Cross-domain on one of the three datasets, and test on the others. The results are presented in Figure 4, which illustrated that TaHiD could better detect other types of fake news even when they are not explicitly used for training. TaHiD achieves the average F1-score of 67.67%, a 28.92% relative improvement compared to Cross-domain. It illustrates that TaHiD learns the information shared by different news propagation networks while previous graph-based methods fail. It further proves that news propagation networks could help to identify fake news.

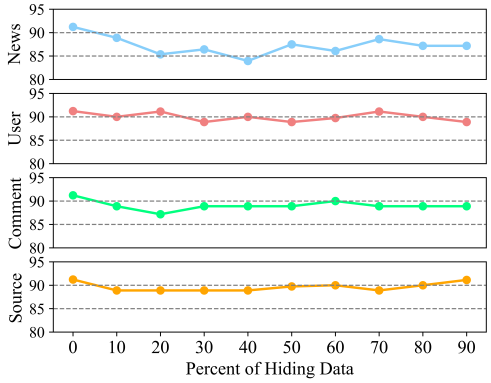


Figure 5: Data hiding study that trains TaHiD on dataset Politifact with hiding data.

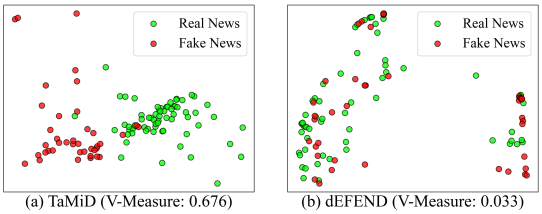
Table 3: The performance of TaHiD without propagation attention module on dataset Politifact. The "Type" column denotes the propagation function of TaHiD.

Type	F1			Acc		
	Prev.	w/o	Diff.	Prev.	w/o	Diff.
GCN	88.8	86.5	-2.3	91.4	90.7	-0.7
HGT	91.0	72.0	-19.0	93.3	80.4	-12.9
GAT	89.2	84.2	-5.0	92.1	88.8	-3.3
R-GCN	89.4	72.2	-17.2	92.1	81.3	-10.8

4.4 TaHiD Data Hiding Study

Data hiding study The ability to solve the data hiding problem means that the model could keep the performance as the news propagation network information disappears. We remove a part of the news propagation network information, which simulates the process of deleting information, and train TaHiD on the removed data. Figure 5 shows the results, which illustrate that TaHiD could keep its performance even if the news itself disappears. In the absence of news itself, TaHiD's F1-score drops by only 7.98%. In the absence of the user, comment, and source information, the F1-score drops by 2.56%, 4.44%, and 2.56% respectively.

Propagation attention study TaHiD contains the propagation module to measure the contribution of different propagations. To prove the ability of this module to identify what matters in the propagation, we train TaHiD without the propagation attention module. Table 3 illustrates the results, which show that the propagation attention module could significantly improve the performance. TaHiD with HGT and R-GCN drop the most, with 20.88% and 19.24% on the F1-score, respectively, which illustrates that the propagation attention module can help significantly improve performance. In other words, TaHiD identifies what matters in the process of news propagation and successfully finds out the hidden fake news.



(a) TaMiD (V-Measure: 0.676) (b) dEFEND (V-Measure: 0.033)

Figure 6: 2D PCA plot of the news representations of TaHiD and dEFEND.

4.5 TaHiD Representation Study

While achieving good performance, TaHiD can also learn the representation of each entity in the news propagation network to support other downstream tasks like political stance detection. We evaluate the representation of news derived from TaHiD compared with dEFEND which also provides news representation. We cluster representations using k-means and calculate the V-Measure (Rosenberg and Hirschberg, 2007), which is an external cluster evaluation. Figure 6 visualizes the representations. Figure 6(a) is the PCD plot of TaHiD representations, which shows moderate collocation for the group of fake and real news, while Figure 6(b) shows little. Quantitatively, TaHiD's clusters achieve a better V-Measure score of 0.676, compared to a 0.033 V-Measure score for the dEFEND clusters.

5 Conclusion

Fake news detection is attracting growing attention in recent years. We propose TaHiD, an end-to-end fake news detection framework that considers news propagation as a heterogeneous dynamic graph. Specifically, TaHiD encodes critical entities in the news propagation including news, source, user, comment, and time information to construct a news propagation network. TaHiD contains a propagation attention model to determine the contribution of different propagation layers to address the data hiding problem. Extensive experiments on three real-world datasets demonstrate that TaHiD achieves excellent performance by considering the news propagation network. Further explorations prove that the propagation attention module is successful and leads to TaHiD's ability to address data hiding problems.

6 Limitations

TaHiD achieves excellent performance on fake news detection and tackles the data hiding problem. TaHiD has two minor limitations:

- TaHiD leverages a news propagation network to identify fake news and achieve excellent performance. In the early days of news propagation, it could not form a relatively large-scale propagation network. Namely, TaHiD's performance in identifying early fake news may drop.
- Extensive experiments show that TaHiD considers the contribution of different propagations. TaHiD cannot give quantitative indicators of the importance of each propagation, which leads to low explainability.

References

Adrien Benamira, Benjamin Devillers, Etienne Lesot, Ayush K Ray, Manal Saadi, and Fragkiskos D Malliaros. 2019. Semi-supervised learning and graph neural networks for fake news detection. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 568–569. IEEE.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Daryna Dementieva and Alexander Panchenko. 2021. Cross-lingual evidence improves monolingual fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 310–320.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shangbin Feng, Zhaoxuan Tan, Rui Li, and Minnan Luo. 2021. Heterogeneity-aware twitter bot detection with relational graph transformers. *arXiv preprint arXiv:2109.02927*.

Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698.

Siva Charan Reddy Gangireddy, Cheng Long, and Tanmoy Chakraborty. 2020. Unsupervised fake news detection: A graph-based approach. In *Proceedings of the 31st ACM conference on hypertext and social media*, pages 75–83.

Yi Han, Shanika Karunasekera, and Christopher Leckie. 2020. Graph neural networks with continual learning for fake news detection from social media. *arXiv preprint arXiv:2007.03316*.

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 (Volume 1: Long Papers)*, pages 754–763.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020*, pages 2704–2710.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. *arXiv preprint arXiv:1903.07389*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Nikhil Mehta and Dan Goldwasser. 2021. Tackling fake news detection by interactively learning representations using graph neural networks. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 46–53.

Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the Web Conference 2021*, pages 3432–3441.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 2173–2178.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

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

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

Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2020b. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019b. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 436–439.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 557–565.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Reuben Tan, Bryan A Plummer, and Kate Saenko. 2020. Detecting cross-modal inconsistency to defend against neural fake news. *arXiv preprint arXiv:2009.07698*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *stat*, 1050:20.

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*, pages 849–857.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 540–547.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural deepfake detection with factual structure of text. *arXiv preprint arXiv:2010.07475*.

Table 4: The user properties and their description that TaHiD adopts.

Metadata	Description
protected	whether this user be protected
followers count	the count of followers
friends count	the count of friends
listed count	the count of lists user follows
activate days	the activate days of user
favourites count	the count of likes the user obtains
geo enable	whether user displays location
verified	whether user is verified
statuses count	the count of statuses
default profile image	whether user uses default image

A Detailed Information of News Propagation Network

The news propagation network contains 4 types of entities including source, news, user, and comment. Each entity has a timestamp identifying the creation time and TaHiD extracts temporal information through the time stamp. TaHiD adopts description to encode source while title and content to encode news. For comment, TaHiD encodes the content information, and adopts the following properties: (i) **reply count**, the count of comments comment this comment; (ii) **favorite count**, the count of users like this comment; (iii) **source**, the platform posting this comment, such as "Twitter Web Client" or "Twitter for Android". TaHiD adopts description and property to encode the user, and the properties TaHiD adopts are shown in Table 4. The news propagation network contains 4 types relations, including (i) **publish**, the source publishes a news; (ii) **discuss**, the news discusses a news; (iii) **post**, the user posts a comment; (iv) **reply**, the comment replies other comments.

B Implement Detailed Information

We submit our code and detailed information as part of the supplementary material[5]. To reproduce our experiment results, we present our hyperparameter setting in Table 5. Our implementation is trained on a GeForce RTX 2080 Ti GPU with 12GB of memory. Under these settings, TaHiD runs a batch for about 60 seconds and the size of the parameters is about 29.5 MB.

[5]Our codes are also available at https://anonymous.4open.science/r/TaHiD-BDA0/

C Extended Experiments Results

To further prove the ability of TaHiD to tackle the data hiding problem, we conduct extended experiments on the other two datasets. Figure 7 and Figure 8 illustrate that FaHiD can keep the performance as the information is absent. In the absence of news, user, comment, and source information, the F1-score drops by 2.52%, 1.73%, 2.04%, and 1.55%, respectively on Gossipcop while F1-score drops by 4.69%, 3.29%, 0.13% and 12.25%, respectively.

Table 5: Hyperparamter setting of TaHiD

Hyperparameter	Value
RoBERTa dim D_s	768
batch size	128
hidden dim D	512
dropout rate	0.5
Optimizer	Adam
learning rate	1e-4
weight decay	1e-5
attention head p	4
num of propagation l	8

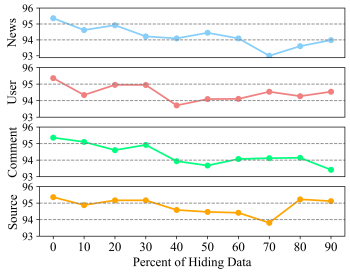


Figure 7: Robustness study that train TaHiD on dataset Gossipcop with hiding data.

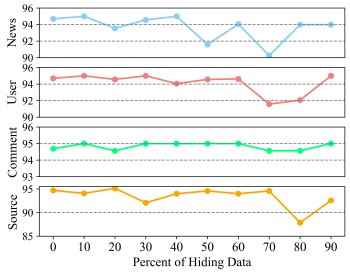


Figure 8: Robustness study that train TaHiD on dataset CoAID with hiding data.