"This is Fake! Shared it by Mistake": Assessing the Intent of Fake News Spreaders

Xinyi Zhou
Syracuse University
Syracuse, NY, USA
zhouxinyi@data.syr.edu

Kai Shu
Illinois Institute of Technology
Chicago, IL, USA
kshu@iit.edu

Vir V. Phoha
Syracuse University
Syracuse, NY, USA
vvphoha@syr.edu

Huan Liu
Arizona State University
Tempe, AZ, USA
huan.liu@asu.edu

Reza Zafarani
Syracuse University
Syracuse, NY, USA
reza@data.syr.edu

ABSTRACT

Individuals can be misled by fake news and spread it unintentionally without knowing it is false. This phenomenon has been frequently observed but has not been investigated. Our aim in this work is to assess the intent of fake news spreaders. To distinguish between intentional versus unintentional spreading, we study the psychological explanations of unintentional spreading. With this foundation, we then propose an *influence graph*, using which we assess the intent of fake news spreaders. Our extensive experiments show that the assessed intent can help significantly differentiate between intentional and unintentional fake news spreaders. Furthermore, the estimated intent can significantly improve the current techniques that detect fake news. To our best knowledge, this is the first work to model individuals' intent in fake news spreading.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Fake news, intent, social media

ACM Reference Format:
Xinyi Zhou, Kai Shu, Vir V. Phoha, Huan Liu, and Reza Zafarani. 2022. "This is Fake! Shared it by Mistake": Assessing the Intent of Fake News Spreaders. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3485447.3512264

1 INTRODUCTION

A frequently observed and discussed phenomenon is that individuals can be misled by fake news and can unintentionally spread it [24, 30, 47]. Thankfully, research has pointed out that (1) correction and (2) nudging can effectively prevent such users from spreading fake news. That is, by informing them of the news falsehood, or simply requesting from them to pay attention to news accuracy before spreading the news [24, 36]. Such findings encourage social media platforms to develop more gentle strategies for these unintentional fake news spreaders to reasonably and effectively combat fake news. Clearly, such strategies should vary from the aggressive deactivation and suspension strategies that platforms adopt for inauthentic or toxic accounts (e.g., Twitter[1] and Facebook[2]). For example, platforms can present such unintentional fake news spreaders with useful facts, motivating the need for new recommendation algorithms. Such algorithms not only recommend topics to these users that they enjoy reading the most (or users they are similar to), but also facts or users active in fact-checking (see Figure 1 for an example) [18, 36, 47].

To determine (1) if correction or nudging is needed for a fake news spreader, (2) whether the spreader should be suspended or deactivated, or (3) which users should be targeted by fact-presenting recommendation algorithms, one needs to assess the *intent* of fake news spreaders. Furthermore, knowing that some users had malicious intent in the past provides a strong signal indicating that their future posts are also potentially fake. This information can be immensely useful for fake news detection [47]. While determining the intent is extremely important, it is yet to be investigated.

This work: Assessing Spreading Intent. We aim to assess the intent of individuals spreading fake news. Our approach to assessing the intent of fake news spreaders relies on fundamental social science theories and exploits advanced machine learning techniques. In particular, we first look into psychological factors that can contribute to the unintentional spreading of fake news (see Section 2.1). These factors can be categorized as *internal influence* and *external influence* [47]. To capture these factors, and in turn, quantify intent, we propose an *influence graph*; a directed, weighted, and attributed graph. The degree to which fake news spreaders are intentional/unintentional can be assessed with this graph. To evaluate our assessment, we first extend two fake news datasets by introducing annotated intent data of fake news spreaders (*intentional* or *unintentional*) due to the unavailability of ground truth.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512264

[1] https://help.twitter.com/en/rules-and-policies/twitter-rules
[2] https://transparency.fb.com/policies/

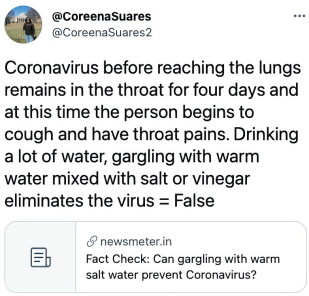


Figure 1: An Example of a Fact-checking Post

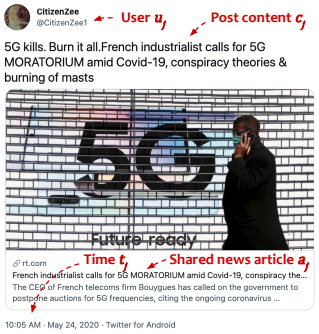


Figure 2: An Illustration of a Post $p_j = (a_j, c_j, t_j, u_j)$

With this data, we validate the assessed intent and show that it can strongly differentiate between intentional and unintentional fake news spreaders. We further show through experiments that the assessed intent can significantly enhance fake news detection.

The innovation and contribution of this work are:

(1) *Modeling Fake News Spreading Intent*: To our best knowledge, this is the first work to assess the degree to which fake news spreaders are intentional/unintentional. To this end, we conduct an interdisciplinary study that endows our work with a theoretical foundation and explainability. A new influence graph is proposed that captures factors that contribute to spreading intent as well as multimodal news information.
(2) *New Datasets on Intent*: We leverage manual and automatic annotation mechanisms to introduce the ground truth on the intent of fake news spreaders in two large-scale real-world datasets. These are the first two datasets that provide intent information. We conduct extensive experiments using these datasets to validate the assessed intent of fake news spreaders.
(3) *Combating Fake News*: Our work helps combat fake news from two perspectives. First, we demonstrate that by assessing intent, we can successfully distinguish between malicious fake news spreaders (should be blocked) and benign ones (should be presented with facts or nudged). Second, we present the effectiveness of the assessed spreader intent (and the proposed influence graph) in fake news detection.

The rest of the paper is organized as follows. A literature review is first conducted in Section 2. In Section 3, we specify the method to assess the intent of fake news spreaders, followed by the method evaluation in Section 4. We demonstrate the value of assessing intent in combating fake news in Section 5. Finally, we conclude in Section 6 with a discussion on our future work.

2 RELATED WORK

We first review fundamental social science theories that have been connected to fake news spreading (see Section 2.1). Next, we review the methods developed to combat fake news (see Section 2.2) as we will later utilize the assessed spreader intent to detect fake news.

2.1 Social Science Foundation of Unintentional Fake News Spreading

Extensive social science research has been conducted on fake news. We particularly review studies that focus on the psychological factors that contribute to the unintentional spreading of fake news.

Lazer et al. [15] attribute this phenomenon to "individuals prefer information that confirms their preexisting attitudes (*selective exposure*), view information consistent with their preexisting beliefs as more persuasive than dissonant information (*confirmation bias*), and are inclined to accept information that pleases them (*desirability bias*)." Scheufele and Krause [30] summarize these factors as *confirmation bias*, *selective exposure*, and *motivated reasoning* (i.e., people tend to use emotionally biased reasoning to make most desired decisions rather than those that accurately reflect the evidence).

Grouping aforementioned psychological factors as an *internal influence*, Zhou and Zafarani [47] further discuss how the *external influence* on individuals can contribute to their unintentional spreading of fake news. Such social influence can be reflected via, e.g., *availability cascade* (i.e., individuals tend to adopt insights expressed by others when such insights are gaining more popularity) [14], *social identity theory* [3, 13] (i.e., individuals conform to the behavior of others for being liked and accepted by the community and society), and *validity effect* (e.g., individuals tend to believe information is correct after repeated exposures) [6, 23].

This work shares the social science foundation presented in [15, 30, 47]. Besides understanding why individuals can be misled by fake news and unintentionally spread it, we further conduct quantitative research to assess user intent.

2.2 Methodologies to Combat Fake News

The unprecedented growth of fake news and its detrimental impacts on democracies, economies, and public health has increased the demand for automatic methodologies to combat fake news [47]. With extensive recent contributions by the research community, automatic fake news detection has significantly improved in efficiency and explainability. In general, fake news detection methods can be content-based or propagation-based depending on whether the method focuses on investigating news content or how the news spreads on social media.

As news articles are mostly text, content-based methods start with manually extracting linguistic features for news representation; LIWC (Linguistic Inquiry and Word Count) [22] has been often employed as a comprehensive feature extractor [7, 25, 27]. Common classifiers, such as SVMs (Support Vector Machines), are then used to predict fake news. With advances in deep learning, recent attention has been paid to employing multimodal (textual and visual) information of news content to detect fake news (see related work such as [1, 26, 39, 43, 46]). On the other hand, propagation-based methods utilize auxiliary social-media information to predict fake news. Some examples of such information include post stances [33], post-repost relationships [37], user comments [32], and profiles [9].

There have been other strategies proposed to combat fake news. For example, education and nudging have been emphasized to improve individuals' ability to recognize misinformation [15, 17, 30]. Pennycook et al. further provide empirical evidence that unintentional fake news spreading can become less by asking individuals

to assess the accuracy of news before attempting to spread it [24]. Lazer et al. suggest incorporating *information quality* into algorithmic rankings or recommendations of online platforms [15]. Studies have also demonstrated that connecting users active in fact checking with fake news spreaders on social networks is an effective way to combat fake news [18, 36].

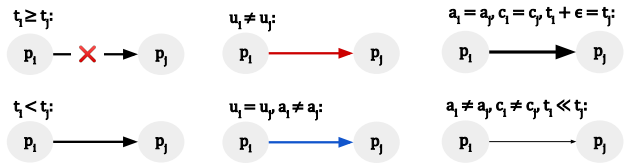
3 MODELING THE INTENT OF FAKE NEWS SPREADERS ON SOCIAL MEDIA

As presented in Section 2.1, psychological factors that contribute to unintentional fake news spreading of individuals can be summarized as two: (1) internal influence and (2) external influence [15, 30, 47]. Hence, an individual is more unintentional in spreading a news article if his or her spreading behavior receives more internal and external influence. Specifically, both *confirmation bias* [20, 21] and *selective exposure* [12, 19] point out that the more consistent an individual's preexisting attitudes and beliefs are with the fake news, the higher the probability that the individual would believe the fake news and unintentionally spread it (internal influence) [15, 30, 47]. As *availability cascade* [14] and *social identity theory* [3, 13] suggest, individuals can be affected by others as well. An individual would be more unintentional in spreading a fake news article if the spreading follows a *herd behavior*; i.e., the individual's participation matches extensive participation of others and his or her attitude conforms to the attitudes of most participants (external influence) [47].

Problems then arise on social media: *where can one find out the preexisting attitudes and beliefs of a user, the participation of users, and their attitudes towards a news article?* We note that a user's preexisting attitudes and beliefs can be reflected in his or her historical activities on social media. For most social media sites, such historical activities include past posts, likes, and comments. Similarly, the participation of users often takes the form of posting, liking, and commenting. Hence, mining the content of posts and comments allows understanding users' attitudes. For simplicity, we start with posts in this work to determine users' preexisting beliefs and participation. In sum, a user spreads a fake news article in his or her post more unintentionally if the post is more similar to or influenced by (1) the user's past posts (internal influence), and (2) the posts of other users (external influence).

A natural approach to capture the influence among posts is to construct an *influence graph* of posts. In this graph, a (directed) edge between two posts indicates the (external or internal) influence flow from one post to the other. The edge weight implies the amount of the influence flow. With this graph, the overall influence that a post receives from other posts can be assessed by looking at its corresponding incoming edges and their weights. The more influence a post that contains fake news receives, the more unintentional is the user who is posting it in spreading this fake news.

To concretely present our proposed influence graph formed by a group of posts, we start by a pair of **p**osts p_i and p_j, which are represented as tuples (a_i, c_i, t_i, u_i) and (a_j, c_j, t_j, u_j), respectively. An example is presented in Figure 2. In the tuple representing post p_i, a_i denotes the news **a**rticle shared by post p_i. For simplicity, we first assume that each post can only share one news article (we will consider a more general case later in this section); **U**ser u_i and **t**ime t_i refer to the user and posting time of p_i; and **C**ontent c_i is



(a) Without (upper fig.) v.s. With (down) Influence (b) External (up) v.s. Internal (down) Influence (c) Large (up) v.s. Small (down) Volume of Influence

Figure 3: Pairwise Influence of Posts p_i and p_j: (a) decides if there is an edge from p_i to p_j in an influence graph; (b) identifies the edge attribute; and (c) determines the edge weight.

the post content, often containing the attitude and opinion of u_i regarding a_i. Next, we discuss how (A) internal and (B) external influence between p_i and p_j can be modeled, respectively.

A. Modeling Internal influence between p_i and p_j. If p_i internally influences p_j, p_i should be posted earlier than p_j and by the same user of p_j (to capture preexisting beliefs of the user), i.e., $t_i < t_j$ and $u_i = u_j$. The amount of influence flowing from p_i to p_j can be determined by how similar the news articles and attitudes in p_j are to those of p_i. In other words, how similar a_j and c_j are to a_i and c_i [47]. However, evidence has indicated that the same user spreading the same news, especially fake news, is often a sign of intentional spreading rather than unintentional spreading [31]. Therefore, we exclude internal influence from p_i and p_j if $a_i = a_j$.

B. Modeling External influence between p_i and p_j. If p_i externally influences p_j, p_i should, at least, be posted by a different user from that of p_j (to capture "external") and earlier than p_j (otherwise, p_i is not observable to p_j); i.e., $t_i < t_j$ and $u_i \neq u_j$. We further consider two questions in assessing external influence. First, can a user's post spreading one news article externally influence a post of another user spreading a different news article; in other words, if $a_i \neq a_j$, can p_i possibly influences p_j externally with $t_i < t_j$ and $u_i \neq u_j$? Two news articles that differ in text or image may discuss the same event and express the same political stance; hence, this scenario is possible but depends on the similarity between the two news articles [5]. Second, can a user's post possibly be influenced by the other's post if the two users are not socially connected on social media? Due to the platforms' diverse recommendations and services (e.g., the trending in Twitter and Weibo), this scenario is also possible, but the amount of influence depends on how similar news articles and attitudes in p_j are to those of p_i [44, 47].

We summarize the above discussions by answering the following three questions:

(1) *Edge existence: Can p_i possibly influence p_j?* As discussed, it is is barely possible for p_i to (internally or externally) affect p_j, if it is posted later than p_j. Hence, in an influence graph, a directed edge can possibly exist from p_i to p_j, if p_i is posted earlier than p_j (i.e., $t_i < t_j$); if $t_i \geq t_j$, no edge exists from p_i to p_j. Therefore, there can be either no edge or only one directed edge between two posts. See Figure 3(a) for an illustration. Note that whether an edge ultimately exists between two posts also depends on the edge weight

(we specify below in 3); a zero weight can make an edge "disappear."

(2) *Edge attribute: What type of influence (internal vs external) is flowing between p_i and p_j?* We define the influence as external, if p_i and p_j are posted by different users, i.e., $u_i \neq u_j$ [44]. The influence is internal, if p_i and p_j are posted by the same user and do not share the same news, i.e., $u_i = u_j$ and $a_i \neq a_j$ [31]. See Figure 3(b) for an illustration.

(3) *Edge weight: How much influence flows from p_i to p_j?* We assume that the amount of influence flow is affected by three factors. The first, as discussed, is the *news articles shared by p_i and p_j (a_i versus a_j)*; basically, if p_i and p_j spread the same news, the influence flow between them should be greater compared to if they spread completely different news articles [5, 47]. The second, as discussed, is the *attitudes held by p_i and p_j on the news (c_i versus c_j)*; basically, if two posts agree with each other, the influence flow between them should be greater compared to if they disagree with each other [47]. Furthermore, we consider the *time interval between p_i and p_j (t_i versus t_j)*; instead of "remembering all", users forget past news articles and their corresponding posts over time (with some decay) [40]. Thus, a greater amount of influence flow is assigned to two posts when one is published close in time to the other, compared to those that are published farther apart.

Next, we formalize the proposed *influence graph* (see Definition 3.1), and introduce how the intent of (fake) news spreaders can be quantified based on this graph. Clearly, in a real-world scenario, it is possible for a post to contain more than one news article (e.g., multiple URLs). Hence, in this formalization, we no longer assume that each post can only share one news article and generalize to a set of articles, i.e., (a_i, c_i, t_i, u_i) becomes (A_i, c_i, t_i, u_i).

Definition 3.1 (Influence Graph). Given a set of news articles, denoted as $A = \{a_1, a_2, \cdots, a_m\}$, we denote user posts that share these news articles on social media as $P = \{p_1, p_2, \cdots, p_n\}$. Each post p_i ($i = 1, 2, \cdots, n$) is represented as a tuple (A_i, c_i, t_i, u_i), where A_i, c_i, t_i, and u_i respectively refer to a set of news articles (can be one article) shared by the post (i.e., $A_i \subseteq A$), the post content, the posting time, and the user.

Influence graph, denoted as $G = (V, E, \mathbf{W})$, is formed by user posts, i.e., $V = P$. Edges exist from p_i to p_j if (i) p_i is posted earlier than p_j, and (ii) p_i and p_j do not share the same news when posted by the same user. In other words, $(p_i, p_j) \in E$ if (i) $t_i < t_j$ and (ii) $A_i \neq A_j$ for $u_i = u_j$. The edge weight for (p_i, p_j) is

$$\mathbf{W}_{ij} = \bar{\mathcal{S}}(A_i, A_j) \cdot \mathcal{S}(c_i, c_j) \cdot \mathcal{T}(\Delta t_{ij}), \tag{1}$$

where $\mathcal{S}(*_i, *_j)$ assesses the similarity between $*_i$ and $*_j$, $\mathcal{T}(\Delta t_{ij})$ for $\Delta t_{ij} = t_j - t_i$ is a self-defined monotonically decreasing decay function to capture users' forgetting, and $\bar{\mathcal{S}}(A_i, A_j)$ computes the average pairwise similarity among news pairs $(a_i, a_j) \in A_i \times A_j$. Formally,

$$\bar{\mathcal{S}}(A_i, A_j) = \frac{\sum_{(a_k, a_l) \in A_i \times A_j} \mathcal{S}(a_k, a_l)}{|A_i| \times |A_j|}; \tag{2}$$

hence, $\bar{\mathcal{S}}(A_i, A_j) = \mathcal{S}(a_k, a_l)$ if $A_i = \{a_k\}$ and $A_j = \{a_l\}$.

Based on the above graph, the overall influence on each post, which we denote as the *affected degree*, is computed as

$$\mathbf{f}_j = \sum_{(p_i, p_j) \in E} \mathbf{W}_{ij}, \tag{3}$$

where the external and internal influence, respectively, refer to

$$\begin{array}{rcll} \mathbf{f}_j^{\text{External}} & = & \sum_{(p_i, p_j) \in E} \mathbf{W}_{ij} & \text{if } u_i \neq u_j; \\ \mathbf{f}_j^{\text{Internal}} & = & \sum_{(p_i, p_j) \in E} \mathbf{W}_{ij} & \text{if } u_i = u_j. \end{array} \tag{4}$$

For posts sharing fake news articles, greater values of $\mathbf{f}_j^{\text{External}}$, $\mathbf{f}_j^{\text{Internal}}$, and $\mathbf{f}_j$ indicate that user j receives more external, internal, and combined (external+internal) influence when spreading the fake news article, i.e., the user engages more unintentionally. Conversely, smaller values of $\mathbf{f}_j^{\text{External}}$, $\mathbf{f}_j^{\text{Internal}}$, and $\mathbf{f}_j$ indicate that the user is affected less and engages more intentionally in fake news spreading.[3]

Customized Implementation Details. The implementation of influence graph has several customizable parts; it can be modified by defining different $\mathcal{T}$, developing different techniques to represent news articles and user posts, and designing ways to compute their similarities. Below are our implementations and justifications.

To represent news articles and posts, we investigate both textual and visual information within the content. Textual features are extracted using transformers, which have excellently performed in understanding semantics of text and various NLP (Natural Language Processing) tasks such as machine translation and sentiment analysis [35, 41]. As user posts are often short and within 512 words (e.g., on Twitter, the number of words are not allowed to exceed 280),[4] we use a pre-trained Sentence-RoBERTa model, which modifies RoBERTa by the Siamese network, to obtain the post embedding [28]; the model performs best in the task of semantic textual similarity.[5] Differently, as news articles are often long and over 512 words,[6] we employ Longformer [4] to derive the semantically meaningful text embedding of news articles. Longformer addresses the limitation of 512 tokens in BERT by reducing the quadratic scaling (with the input sequence) to linear [4]. For visual features, we extract them using a pre-trained DeepRanking model particularly designed for the task of fine-grained image similarity computation [38]. With textual features of news (or post) denoted as $\mathbf{t}$, and its visual features denoted as $\mathbf{v}$, we define the similarity between a news (or post) pair as

$$\mathcal{S}(*_i, *_j) = \mu \hat{\cos}(\mathbf{t}_{*_i}, \mathbf{t}_{*_j}) + (1 - \mu) \hat{\cos}(\mathbf{v}_{*_i}, \mathbf{v}_{*_j}), \tag{5}$$

where $* = a$ (for news) or p (for posts); $\hat{\cos}(.,.) = [1 - \cos(.,.)]/2$; and μ, $\hat{\cos}(.,.)$, $\mathcal{S}(.,.) \in [0, 1]$. In our experiments, we determine the value of μ by varying it from 0.1 to 0.9 with a step size 0.1; we set $\mu = 0.8$ that leads to the best evaluation and prediction results.

As for decay function $\mathcal{T}$, we define it as

$$\mathcal{T}(\Delta t_{ij}) = e^{1 - \Delta t_{ij}}, \tag{6}$$

which is inspired by [40]. $\Delta t_{ij} = t_j - t_i$ and t_i indicates the chronological ranking of post p_i (i.e., $t_i \in \mathbb{Z}^+$); hence, $\mathcal{T}(.) \in (0, 1]$ due

[3] The statement also holds for posts sharing true news articles.
[4] https://developer.twitter.com/en/docs/counting-characters
[5] https://github.com/UKPLab/sentence-transformers
[6] As [45] suggests, the number of words of news articles published by mainstream and fake news medium has a mean value around 800 and median value around 600.

to $t_j > t_i$. The benefit of such $\mathcal{T}$ is two fold. First, it helps normalize the affected degree for any influence graph. Specifically, let $\mathbf{f}_j^*$ denote either of $\mathbf{f}_j$, $\mathbf{f}_j^{\text{INTERNAL}}$, or $\mathbf{f}_j^{\text{EXTERNAL}}$. Let $\hat{\mathbf{f}}_j^*$ denote the normalized version of $\mathbf{f}_j^*$, i.e., $\hat{\mathbf{f}}_j^* \in [0, 1]$ (accurately, here $\hat{\mathbf{f}}_j^* \in [0, 1)$). Then, for $\mathbf{f}_j^*$ we have

$$\begin{aligned} \mathbf{f}_j^* &= \sum_{(p_i,p_j)\in E} \mathcal{S}(A_i, A_j) \cdot \mathcal{S}(c_i, c_j) \cdot \mathcal{T}(\Delta t_{ij}) \\ &\leq \sum_{(p_i,p_j)\in E} \mathcal{T}(\Delta t_{ij}) \\ &< \sum_{k=1}^{\infty} e^{1-k} \\ &= e(e-1)^{-1}. \end{aligned} \tag{7}$$

In other words, the upper bound of the affected degree, denoted by $f_{\max}$, is $e(e-1)^{-1}$. Strictly speaking, K posts ($K > 1$) can be posted at the same time in a real-world scenario, i.e., their ranking, denoted by t_X, is the same. We point out that the upper bound $f_{\max}$ still holds in this case, if the ranking value after t_X is $t_X + K$ rather than $t_X + 1$. Finally, the normalized affected degree $\hat{\mathbf{f}}_j^*$ for post p_j is

$$\hat{\mathbf{f}}_j^* = \frac{1}{f_{\max}} \mathbf{f}_j^* = \frac{e-1}{e} \mathbf{f}_j^*. \tag{8}$$

Secondly, in the worst case, influence graph can be a *tournament*, taking up much space. Such $\mathcal{T}$ facilitates *graph sparsification*, while maintaining the performance on tasks (see details in Appendix A). Lastly, we note that we have tested Δt_{ij} (the time interval) with various units (seconds/minutes/hours/days) in addition to chronological rankings; still, the ranking performs best in all experiments.

4 METHOD EVALUATION

In this section, we evaluate the proposed method in assessing the intent of fake news spreaders. To this end, evaluation data is required that contains the ground-truth label on

- *News credibility*, i.e., whether a news article is fake news or true news; and
- *Spreader intent*, i.e., whether a user spreads a fake news article intentionally or unintentionally on social media.

We point out that this work is the first to model individuals' intent in fake news propagation. Therefore, no data exists that contains the ground-truth label on spreader intent, let alone both news credibility and spreader intent. Next, we first detail how this problem is addressed in Section 4.1, followed by the method evaluation results in Section 4.2.

4.1 Datasets and Annotations

Our experiments to evaluate the proposed method are based on two datasets developed for news credibility research: MM-COVID [16] and ReCOVery [45]. Generally speaking, both datasets collect news information verified by domain experts (labeled as *true* or *fake*) and how the news spreads on Twitter. The corresponding data statistics are in Table 1(a); we focus on the news with social context information, and on the English news and tweets to which all pretrained models can be applied.

Although the ground-truth label on news credibility is available, both datasets do not provide annotations on intent of fake news spreaders. We first consider *manual annotation* to address this problem. Specifically, we invite one expert knowledgeable in misinformation area and one graduate student generally aware of the area. We randomly sample 300 posts (unique in tweet ID and user ID) from MM-COVID and ReCOVery that contain fake news (i.e., users of these posts are all fake news spreaders). Before annotating, we first inform the annotators with the definition and general characteristics of unintentional fake news spreaders. That is, as presented in Section 1: these spreaders are misled by fake news, barely recognize it is fake, tend to believe in the fake news; meanwhile, if informed on news falsehood or presented with facts, such spreading behavior of them can be reduced, or even stopped. In annotating, we present the two annotators with

- The tweet's link that spreads fake news, which allows annotators to access the tweet details (as illustrated in Fig. 2).
- The user's link who posts the tweet, which allows annotators to access the user's profile and historical activities.

For each post, we ask the two annotators to

(1) Annotate if the user spreads the fake news unintentionally (with an answer of *yes* or *no*);
(2) Present the confidence level (detailed below);
(3) Explain the annotation with evidence; and
(4) Provide an estimate on the time spent on annotation.

We provide three optional levels of confidence. 0 indicates the annotation result is a random guess; no evidence is found to help annotation, or half the evidence supports but the other half rejects the annotation result. 0.5 indicates a medium-level confidence; among all the evidence that the annotator finds, some of them reject but most of them support the annotation result. 1 indicates a high-level confidence; all the evidence that the annotator finds support the annotation result.

With the returned annotations, we compute the agreement of the two annotators by Cohen's κ coefficient [10]. $\kappa = 0.61$, removing annotations with no confidence; in other words, two annotators substantially agree with each other [10]. To further obtain the ground truth, we only consider the annotations with a confidence score ≥ 0.5 and agreed by the two annotators. Finally, 119 posts sharing fake news have the ground-truth label on their users' intent, among which 59 are unintentional and 60 are intentional.

We point out that annotating intent of fake news spreaders is a time-consuming and challenging task. Around five minutes is required to annotate each instance on average. Understanding the user intent behind a post demands evaluating the tweet content and studying the user based on his or her historical behavior on social media. Such manual annotation for large-scale data is hence impractical, which drives us to consider *algorithmic annotation* that accurately simulates manual annotation in an automatic manner. Interestingly, we observe that annotators are more confident in identifying intentional fake news spreaders than unintentional ones. Specifically, the expert annotator is at 0.93 confidence level in identifying intentional fake news spreaders and at 0.75 confidence level in identifying unintentional fake news spreaders. For the graduate student annotator, the confidence score is 0.84 and 0.57, respectively. Both results have $p \ll 0.001$ with Mann–Whitney U test. To conduct algorithmic annotation that can accurately simulate manual annotation, we thus start to think *"what kind of fake news spreaders can be intentional?"*

Table 1: Data Statistics

(a) on News Credibility

		MM-COVID	ReCOVery
# News	Fake	355	535
	True	448	1,231
# Tweets	Sharing Fake News	16,500	26,657
	Sharing True News	20,905	117,087

(b) on Intent of Fake News Spreaders

		MM-COVID	ReCOVery
# Fake News Spreaders	Unintentional	9,237	7,911
	Intentional	4,285	7,327
	Bots	3,195	6,266
	Trolls	1,024	2,687
	Correctors	463	6
# Tweets Sharing Fake News	by Unintentional	10,519	10,733
	by Intentional	5,953	12,502
	by Bots	4,530	11,035
	by Trolls	1,360	4,240
	by Correctors	789	8

With the explanations given by annotators, we can reasonably assume bots and trolls who have engaged in fake news propagation as intentional fake news spreaders. As inauthentic and toxic accounts, bots and trolls have been often suspended or deactivated by social media platforms (e.g., Twitter and Facebook) regardless of spreading fake news or not. In fact, they have played a significant role in fake news dissemination [11, 31, 34, 47]. As a comparison, unintentional fake news spreaders deserve a "gentle" strategy developed by social media platforms: nudging and fact-presenting recommendation are more reasonable than suspension and deactivation, as we specified in Section 1. Therefore, we separate bots and trolls from unintentional fake news spreaders. We further notice that users active in fact-checking can spread fake news as well, in a *correction* manner; i.e., they clarify news is false (objectively, and not aggressively) and inform other users of it in their spreading. We call the corresponding posts that spread fake news *correction posts* and these users *correctors* later in the paper. These correctors enables recognizing news falsehood. We thus separate them from unintentional fake news spreaders.

We identify bots and trolls by collecting data from two well-established and widely accepted platforms, Botometer [29][7] and Bot Sentinel.[8] Ultimately, each Twitter user is assigned a bot score (denoted as b) and a troll score (denoted as r), where $b, c \in [0, 1]$. To identify correctors, we first annotate each tweet as a correction or non-correction tweet. Then, we assign each fake news spreader a corrector score (denoted as c, where $c \in [0, 1]$) by computing the proportion of the user's correction tweets to his or her total tweets that share fake news. With a threshold value, $\theta \in [0, 1]$, each fake news spreader can be classified as (i) bot (if $b \in [0, \theta)$) or non-bot (if $b \in [\theta, 1]$), (ii) troll (if $r \in [0, \theta)$) or non-troll (if $r \in [\theta, 1]$), and (iii) corrector (if $c \in [0, \theta)$) or non-corrector (if $c \in [\theta, 1]$).

[7] https://botometer.osome.iu.edu/
[8] https://botsentinel.com/

Table 2: Performance of Algorithmic Annotations on Intent of Fake News Spreaders

	AUC Score	Cohen's κ
MM-COVID + ReCOVery	0.8824	0.7482
MM-COVID	0.8857	0.7520
ReCOVery	0.8000	0.6484

With identified bots, trolls, and correctors (here, we use 0.5 as the threshold, i.e., $\theta = 0.5$), the algorithmic annotation on intent of fake news spreaders is conducted at two levels: (i) tweet-level and (ii) user-level. At the tweet-level, the algorithm labels all correction tweets and tweets of bots and trolls that share fake news as intentional spreading. The tweet-level annotation captures the user intent for *each* spreading action of fake news. At the user-level, the algorithm labels all bots, trolls, and correctors as intentional spreaders. The user-level annotation captures the *general* user intent when spreading fake news. Table 1(b) summarizes the corresponding data statistics.

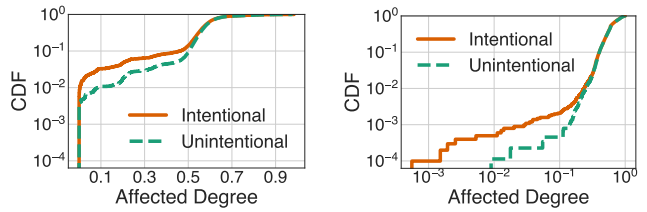
Evaluating Algorithmic Annotations. We compare the algorithmic annotation results with the manual annotations. Results are shown in Table 2; results are the same at both the tweet- and user-levels. We observe that the algorithmic annotation effectively simulates the manual annotation, whose AUC score is above 0.8 using sampled MM-COVID and/or ReCOVery datasets. Automatic and manual annotations have a substantial agreement with Cohen's κ coefficient above 0.64 [10].

4.2 Experimental Results

With annotated intent (*intentional* or *unintentional*) of fake news spreaders, we verify if the assessed intent (i.e., affected degree) differs between intentional and unintentional fake news spreaders and if such difference is statistically significant. In particular, our assessed intent can be validated if affected degrees of intentional fake news spreaders are significantly less than that of unintentional fake news spreaders, i.e., if we estimate fake news spreaders who are annotated as unintentional to be more unintentional than those who are annotated as intentional.

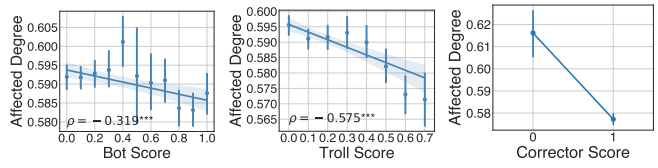
As specified in last section, annotations are conducted at both tweet and user levels. Correspondingly, affected degrees are computed at two levels; we further obtain the user-level affected degree by averaging the affected degree of the user's posts sharing fake news. Here we present tweet-level verification results; results at the two levels reveal the same pattern, from which we can draw the same conclusions.

First, we present the distribution of affected degrees for intentional and unintentional fake news spreaders (see Figure 4). We observe that, in general, the affected degree of intentional fake news spreaders is less than that of unintentional fake news spreaders. Specifically, the average *normalized* affected degree of intentional fake news spreaders are 0.55 with MM-COVID data and 0.61 with

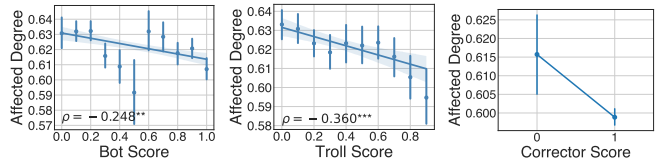


(a) MM-COVID ($p \ll 0.001$ with t-test) (b) ReCOVery ($p < 0.01$ with t-test)

Figure 4: Distribution of Affected Degree: Intentional Fake News Spreaders v.s. Unintentional Fake News Spreaders

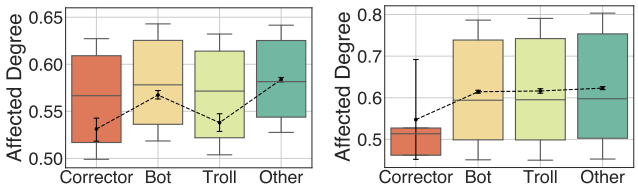


(a) MM-COVID ($p \ll 0.001$ using t-test for the right)



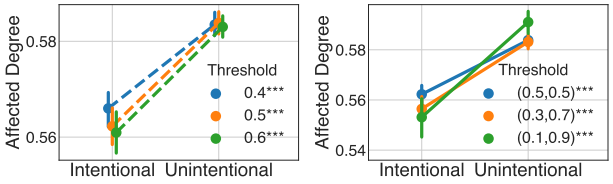
(b) ReCOVery ($p \ll 0.001$ using t-test for the right)

Figure 6: Relation between Affected Degree and (L) Bot Score, (M) Troll Score, and (R) Corrector Score. ρ: Spearman's Correlation Coefficient. ***: $p < 0.001$; **: $p < 0.01$; and *: $p < 0.05$.

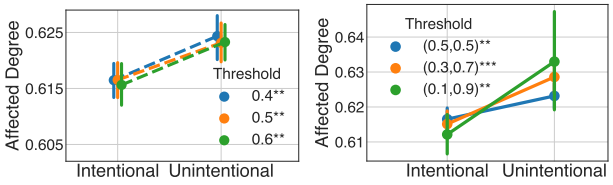


(a) MM-COVID ($p \ll 0.001$ by ANOVA) (b) ReCOVery ($p < 0.01$ by ANOVA)

Figure 5: Affected Degree of Bots, Trolls, Correctors, and Others (First Three: Intentional Fake News Spreaders; Others: Unintentional Fake News Spreaders)



(a) MM-COVID



(b) ReCOVery

Figure 7: Method Performance with Various Thresholds (*: $p < 0.001$; **: $p < 0.01$; and *: $p < 0.05$)**

ReCOVery data. For unintentional fake news spreaders, the value is 0.58 and 0.62, respectively. Such difference is statistically significant with a p-value of $\ll 0.001$ on MM-COVID and < 0.01 on ReCOVery using t-test. Therefore, the results validate our assessment. We conduct the same experiment on the subset of data annotated by humans, where we can draw the same conclusion.

Second, we compare the affected degree of bots, trolls, and correctors, which all are annotated as intentional fake news spreaders, with that of others, which are annotated as unintentional fake news spreaders. The results are shown in Figure 5. The results indicate that bots, trolls, and correctors all have a lower affected degree compared to unintentional fake news spreaders. The results are statistically significant with a p-value of $\ll 0.001$ on MM-COVID and < 0.01 on ReCOVery using ANOVA test. Meanwhile, Figure 6 presents the relationship between affected degree and (i) bot score, (ii) troll score, and (iii) corrector score. The results reveal the same pattern: affected degree drops with an increasing bot, troll, or corrector score. In particular, both bot and troll scores are negatively correlated with affected degrees, with a Spearman's correlation coefficient $\rho \in [-0.32, -0.24]$ for bots and $\rho \in [-0.58, -0.36]$ for trolls. Results, again, validate our proposed method. Note that when investigating the relationship between affected degree and, e.g., bot score, we remove trolls and correctors to reduce noise.

Third, we assess the result robustness. As mentioned before, a fake news spreader is labeled as an unintentional spreader with a bot (troll, or corrector) score less than a threshold value (i.e., $X \in [0, \theta), X = \{b, r, c\}$); otherwise, he or she is an intentional spreader (i.e., $X \in [\theta, 1], X = \{b, r, c\}$). Varying θ among $0.4, 0.5, 0.6$, we compare again the affected degree of intentional and unintentional fake news spreaders. Results are presented in Figure 7 (the left column). We observe that slightly adjusting the threshold value does not change our observations and conclusions made in the first experiment (i.e., the result is robust).

We lastly evaluate the proposed method as follows: we label a fake news spreader whose $X \in [0, \theta)$ as an unintentional spreader, and whose $X \in [1 - \theta, 1]$ as an intentional spreader. By decreasing θ_X, a fake news spreader is required to have a lower bot (troll, or corrector) score to be unintentional and a higher bot (troll, or corrector) score to be intentional. In other words, a smaller θ corresponds to a more strict annotation (*intentional* or *unintentional*) of fake news spreaders. We vary θ among $0.5, 0.3, 0.1$ – correspondingly, $1 - \theta$ varies among $0.5, 0.7, 0.9$ – and compare the affected degree of

Table 3: Method Performance with Hand-crafted Features in Fake News Detection. Here, K: the first (earliest) K posts spreading the news available for news representation; Ranking: feature importance ranking of affected degree of posts in the prediction model.

	K	AUC Score	Ranking
MM-COVID	10	0.918 (±0.009)	2
	20	0.912 (±0.015)	2
	30	0.927 (±0.021)	2
	40	0.923 (±0.012)	2
	All	0.935 (±0.005)	3
ReCOVery	10	0.891 (±0.007)	5
	20	0.898 (±0.007)	3
	30	0.903 (±0.004)	3
	40	0.909 (±0.014)	4
	All	0.925 (±0.009)	5

intentional and unintentional fake news spreaders. Results are presented in Figure 7 (the right column). We observe that the affected degree of intentional fake news spreaders is always less than that of unintentional fake news spreaders with various thresholds. More importantly, such pattern becomes more significant with a smaller θ (i.e., a more strict annotation), which validates the effectiveness of our assessment.

Finally, we point out that we experiment with (i) external affected degree, (ii) internal affected degree, (iii) combined (external+internal) affected degree, and (iv) combined affected degree where the external one merely exists between post pairs sharing the same news. The combined one (i.e., iii) is the one where significant and consistent patterns are discovered on both datasets.

5 UTILIZING INTENT OF NEWS SPREADERS TO COMBAT FAKE NEWS

Using MM-COVID and ReCOVery data, we evaluate the effectiveness of user intent in news propagation to detect fake news. We first employ the assessed affected degree of posts in news propagation within a traditional machine learning framework. Then, we utilize the proposed influence graph within a deep learning framework.

I. Combating Fake News by Affected Degree. For each news article, we manually extract over 100 (propagation and content) features as its representation. Propagation features include the average (internal, external, and combined) affected degree of posts spreading the news and a set of widely-accepted propagation features. Content features are extracted using LIWC [22]. See Appendix B for feature details. Five-fold cross-validation and XGBoost [8] are then used with these features for training and classifying news articles. Results indicate that this method correctly identifies fake news with an AUC score of around 0.93. As a comparison, dEFEND [32], a state-of-the-art method that detects fake news by news content and propagation information, performs around 0.90. Furthermore, we observe that, as presented in Table 3, the proposed method performs above 0.89 with limited propagation information of news articles, i.e., at an early stage of news dissemination on social media. Notably, internal affected degree of posts greatly contributes to detecting fake news, whose feature importance assessed by XGBoost ranks top five all along.

II. Combating Fake News by Influence Graph. We construct the news-post heterogeneous graph (shown in Figure 8); a post is connected with a news article if the post shares the news, and the relation among posts is modeled by the proposed influence graph G. Then, we train the HetGNN (Heterogeneous Graph Neural Network) model [42] with this news-post graph to learn news representation, with which XBGoost [8] is further utilized to predict fake news. Varying the percentage of labeled news from 20% to 80%, this method performs with an AUC score ranging from 0.83 (with small-scale training data) to 0.91 (with relatively large-scale training data) on two datasets. To further evaluate the proposed influence graph G, we consider two variant groups of the constructed heterogeneous graph as baselines. One replaces G by a random version (G_{RANDOM}): Based on our graph sparsification strategy (see Appendix A), we construct the random graph by randomly selecting a hundred posts for each post ensuring that no self-loops are formed in this graph. The other replaces G by its subgraph (i) with internal influence only (G_{INTERNAL}); (ii) with external influence only (G_{EXTERNAL}); or (iii) with internal and external influence but the latter only exists between two posts sharing the same news ($G_{\text{SAME NEWS}}$). Table 4 presents the full result; G_{SUBGRAPH} in the table refers to $G_{\text{SAME NEWS}}$, which performs best among all subgraphs. We observe that in general, the proposed influence graph outperforms its variants in detecting fake news, especially with limited training data. See Appendix B for other implementation details.

6 CONCLUSION AND FUTURE WORK

We look into the phenomenon that social media users can spread fake news unintentionally. With social science foundations, we propose influence graph, with which we assess the degree to which fake news spreaders are unintentional (denoted as *affected degree*). Strategies to sparse the influence graph and normalize the affected degree by determining its upper bound are presented as well. We develop manual and automatic annotation mechanisms to obtain the ground-truth intent (*intentional* or *unintentional*) of fake news spreaders for MM-COVID and ReCOVery data. We observe that the affected degree of intentional fake news spreaders are significantly less than that of unintentional ones, which validates our assessments. This work helps combat fake news from two perspectives. First, our assessed intent helps determine the necessity of a fake news spreader being nudged or recommended with (users active in sharing) facts. Second, we present that the assessed spreader intent and proposed influence graph effectively help detect fake news with an AUC score of around 0.9.

Limitations and Future Work: We effectively assess the degree to which fake news spreaders are unintentional, but remain the task to *classify* a fake news spreader as an intentional or unintentional spreader. We point out that merely relying on determining a threshold for affected degree is barely enough. To address this problem, we aim to propose a more complicated classification model in the near future, which involves non-posting behavior (e.g., commenting, liking, and following) of news spreaders.

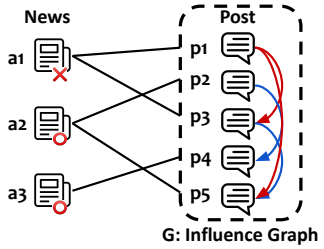


Figure 8: News-post Graph

Table 4: Method Performance (Using AUC Scores) with Heterogeneous Graph Neural Networks (HetGNN) in Fake News Detection

	MM-COVID				ReCOVery			
% Labeled News	20%	40%	60%	80%	20%	40%	60%	80%
G_{RANDOM}	0.829	0.856	0.876	0.902	0.647	0.654	0.660	0.674
G_{SUBGRAPH}	0.817	0.861	0.890	**0.915**	0.820	0.845	0.869	**0.908**
G	**0.869**	**0.864**	**0.902**	0.905	**0.825**	**0.863**	**0.883**	0.881

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under award CAREER IIS-1942929. We sincerely appreciate the positive and constructive comments of the reviewers. We also thank Chang Liu, Shengmin Jin, and Hao Tian for their useful suggestions in data annotation.

REFERENCES

[1] Anton Abilov, Yiqing Hua, Hana Matatov, Ofra Amir, and Mor Naaman. 2021. VoterFraud2020: A Multi-modal Dataset of Election Fraud Claims on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 901–912.

[2] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.

[3] Blake E Ashforth and Fred Mael. 1989. Social Identity Theory and the Organization. *Academy of Management Review* 14, 1 (1989), 20–39.

[4] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150* (2020).

[5] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Viral Misinformation: The Role of Homophily and Polarization. In *Proceedings of the 24th International Conference on World Wide Web*. 355–356.

[6] Lawrence E Boehm. 1994. The Validity Effect: A Search for Mediating Variables. *Personality and Social Psychology Bulletin* 20, 3 (1994), 285–293.

[7] Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A Topic-Agnostic Approach for Identifying Fake News Pages. In *Companion Proceedings of the 2019 World Wide Web Conference*. ACM, 975–980.

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.

[9] Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2021. Causal Understanding of Fake News Dissemination on Social Media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 148–157.

[10] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[11] Emilio Ferrara. 2020. What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday* 6 (2020). https://doi.org/10.5210/fm.v25i6.10633

[12] Jonathan L Freedman and David O Sears. 1965. Selective exposure. In *Advances in Experimental Social Psychology*. Vol. 2. Elsevier, 57–97.

[13] Michael A Hogg. 2020. *Social Identity Theory*. Stanford University Press.

[14] Timur Kuran and Cass R Sunstein. 1999. Availability Cascades and Risk Regulation. *Stanford Law Review* 51, 4 (1999), 683–768.

[15] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[16] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A Multilingual and Multimodal Data Repository for Combating COVID-19 Disinformation. arXiv:2011.04088 [cs.SI]

[17] Philipp Lorenz-Spreen, Stephan Lewandowsky, Cass R Sunstein, and Ralph Hertwig. 2020. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour* 4, 11 (2020), 1102–1109.

[18] Drew B Margolin, Aniko Hannak, and Ingmar Weber. 2018. Political fact-checking on Twitter: When do corrections have an effect? *Political Communication* 35, 2 (2018), 196–219.

[19] Miriam J Metzger, Ethan H Hartsell, and Andrew J Flanagin. 2020. Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research* 47, 1 (2020), 3–28.

[20] Sendhil Mullainathan and Andrei Shleifer. 2005. The market for news. *American Economic Review* 95, 4 (2005), 1031–1053.

[21] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2 (1998), 175–220.

[22] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[23] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General* 147, 12 (2018), 1865.

[24] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

[25] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3391–3401.

[26] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 153–162.

[27] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2931–2937.

[28] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3973–3983.

[29] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2725–2732.

[30] Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7662–7669.

[31] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature Communications* 9, 1 (2018), 4787.

[32] 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.

[33] Panayiotis Smeros, Carlos Castillo, and Karl Aberer. 2019. SciLens: Evaluating the Quality of Scientific News Articles Using Social Media and Scientific Literature Indicators. In *Proceedings of the International Conference on World Wide Web*. ACM, 1747–1758.

[34] Kate Starbird. 2019. Disinformation's spread: bots, trolls and all of us. *Nature* 571, 7766 (2019), 449–450.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 5998–6008.

[36] Nguyen Vo and Kyumin Lee. 2018. The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 275–284.

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

[38] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity

with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1386–1393.
[39] 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.
[40] Piotr A Woźniak, Edward J Gorzelańczyk, and Janusz A Murakowski. 1995. Two components of long-term memory. *Acta Neurobiologiae Experimentalis* 55, 4 (1995), 301–305.
[41] Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3695–3706.
[42] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 793–803.
[43] Jiaxuan Zhang, Sarah Ita Levitan, and Julia Hirschberg. 2020. Multimodal Deception Detection Using Automatically Extracted Acoustic, Visual, and Lexical Features. In *INTERSPEECH*. 359–363.
[44] Xin Zhang, Ding-Ding Han, Ruiqi Yang, and Ziqiao Zhang. 2017. Users' participation and social influence during information spreading on Twitter. *PloS one* 12, 9 (2017), e0183290.
[45] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3205–3212.
[46] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. In *Advances in Knowledge Discovery and Data Mining*, Vol. 12085. Nature Publishing Group, 354.
[47] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.

A SPARSIFICATION OF INFLUENCE GRAPH

Influence graph can be a *tournament* in the worst case, taking much space. To sparsify the graph, we add one more constraint in the graph construction: $(p_i, p_j) \in E$ if $\Delta t_{ij} \leq \theta_t$. Thus, we assume that each node (post) can be connected with (affected by) at most θ_t previous nodes (posts), which can be viewed as an extension of the Markov property. We vary θ_t in {1, 10, 100, 1000} and ultimately set $\theta_t = 100$ as all experimental results converge at this point.

B REPRODUCIBILITY DETAILS IN FAKE NEWS DETECTION

We have 109 hand-crafted (linguistic and propagation) features. Propagation features include the average external, internal, and combined affected degree of posts sharing the news; the average sentiment score (assessed by flair [2])[9]and the average number of reposts, favorites, hashtags, mentions, symbols, quotes, and replies of posts sharing the news; and the average number of followers, friends, favorites, list memberships, and status updates of users spreading the news. Content features include all that can be extracted by LIWC [22], each of which falls into one of the categories including word count, summary language variables, linguistic dimensions, other grammars, and psychological processes.

With HetGNN, we use pre-trained transformers to extract content features of nodes (Longformer [4] for news stories and Sentence-BERT [28] for tweets). The news node is associated with the news embedding and the average embedding of its connected posts. The post node is associated with the post embedding, the average embedding of its connected news, and the average embedding of its connected posts. Hence, the Bi-LSTM length of news content encoder is two, and that of post content encoder is three. For both datasets, the embedding dimension of HetGNN is 1024, the size of sampled neighbors set for each node is 23 (3 news nodes plus 20 post nodes), the learning rate is 0.0001, and the maximum number of training iterations is 50. The other hyperparameters are set the same as mentioned in [42].

[9]https://github.com/flairNLP/flair