An Interactive Framework for Profiling News Media Sources

Nikhil Mehta
Department of Computer Science
Purdue University
West Lafayette, IN 47907
mehta52@purdue.edu

Dan Goldwasser
Department of Computer Science
Purdue University
West Lafayette, IN 47907
dgoldwas@purdue.edu

Abstract

The recent rise of social media has led to the spread of large amounts of fake and biased news, content published with the intent to sway beliefs. While detecting and profiling the sources that spread this news is important to maintain a healthy society, it is challenging for automated systems.

In this paper, we propose an interactive framework for news media profiling. It combines the strengths of graph based news media profiling models, Pre-trained Large Language Models, and human insight, to characterize the social context on social media. Experimental results show that with as little as 5 human interactions, our framework can rapidly detect fake and biased news media, even in the most challenging settings of emerging news events, where test data is unseen.

1 Introduction

The recent rise of social media has enabled information to spread at a rapid pace, having the potential to very quickly impact a large number of people, especially during key events, such as political elections (Vosoughi et al., 2018). While this rise of social media has many benefits, one downside is that harmful information, i.e. fake or politically biased news, can also spread rapidly, affecting people's perspectives. Thus, detecting it is important.

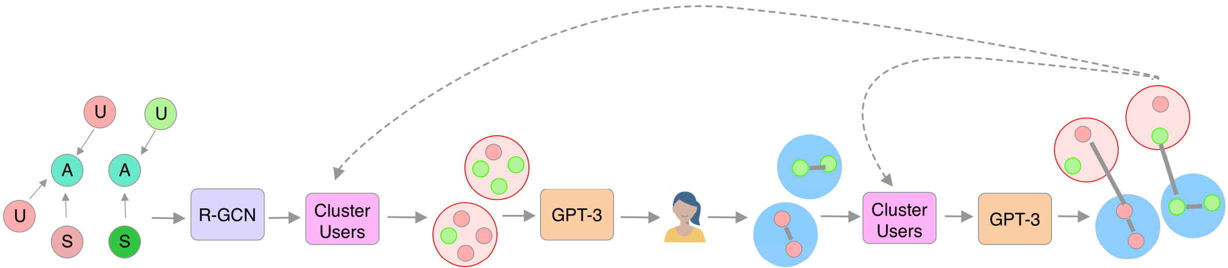
While one approach is to fact-check or detect the bias of all content on social media (i.e. Twitter), another is to focus on the source, and ask: *Is this source factual or politically biased?* This task, **profiling news media sources**, which we focus on, can scale better, as often times a large amount of the content sources' publish have the same factuality/political bias as the source itself. We model this on a 3-point scale: *high, low, and mixed* factuality, and *left, center, and right* bias. Details: App. A.1.

Even at the source level, it is difficult for humans to profile all news content, due to the large number of sources online. Further, this task is still challenging for AI systems (Baly et al., 2018, 2020; Mehta et al., 2022), especially in the *emerging events settings*, when the system is tested on its **ability to adapt to new events**, consisting of new sources, content they generate, and social media users engaging with them that were not seen at training time (Yuan et al., 2020). For example, in a graph framework, test set nodes are not connected to training set nodes. In these settings, Large Language Models (LLMs) also struggle, even when enhanced with extra knowledge (Whitehouse et al., 2022).

Due to the struggles of AI systems to automatically profile news media, in this paper we propose a different, interactive approach, for this task. We are inspired by recent results (Cinelli et al., 2021; Vicario et al., 2016) showing that misinformation and highly biased content tends to spread in closely knit communities on social media. This leads us to ask, whether better modeling of the social media relationships that underlie content spread would improve our ability to profile the content itself. Specifically, we hypothesize that users on social media form *information communities*, or groups of users, where certain themes circulate more in some communities vs. others. If we can identify these themes and use them to **identify information communities**, then we can better profile the content discussed by the communities. For example, a user joining a left-leaning community, is more likely to be left-leaning, and so is the content they share. Then, if that user shares content from a source, that source is also more likely to be left-leaning.

In the settings of emerging events, we must be able to characterize and form the information communities quickly, without using labeled training data, so we can rapidly detect fake/biased news sources on unseen data. Unfortunately, using only information on social media to form information communities involves complex reasoning on unseen data, and is thus challenging for automated

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 40–58
June 16-21, 2024 ©2024 Association for Computational Linguistics



(a) Broad Event Graph (b) Graph Model (c) Clustering to form candidate information communities (d) Using GPT-3 and humans, narrow down communities based on user similarity, then humans connect users in the same community to form validated ones (light blue). (e) Validated Communities are iteratively expanded, by clustering to form new candidates (light red) and assigning their GPT-3 validated users to validated communities (light blue)

Figure 1: Our framework overview: **Using Trained Graph Models, Large Language Models (GPT-3) and Human Interaction to Form Information Communities for News Media Profiling**. (Key: U = Users, A = Articles, S = Sources, Light Red Background = Candidate information Communities, Light Blue Background = Validated Information Communities). From the learned graph model (b), we find candidate information communities through k-means clustering. Using a LLM, GPT-3, we form a textual representation of the information community by summarizing its users, and then ask humans to narrow down the community based on the user summaries (d), forming smaller, validated communities, whose users are then connected to each other. We then expand the validated communities, by again model clustering users, forming user summaries, but this time asking GPT-3 to place or not palce the users into validated communities, which can be done repeatedly (e). This entire process (c-e) can repeat, starting with clustering of unassigned users (c), to form more validated communities, which can be expanded further.

systems, even LLMs (as we experimentally show). Thus, instead, in this paper, we propose an **interactive learning framework**, to form information communities. We take advantage of minimal human interactions, combining it with the strengths of news media profiling graph models and LLMs, to form the communities. We do this in an iterative process, showing how less than 5 human interactions, which do not label any additional data, can lead to significant improvements in both fake news and bias source detection, even in the challenging weakly supervised emerging events settings.

Specifically, to form the communities, we first form initial candidate communities by clustering social media users based on their graph model embeddings, capturing model beliefs. As news media profiling is difficult, these candidate information communities are likely to be imperfect, i.e., some users are likely to be inconsistent with their community assignment. Thus, these assignments need to be validated, by examining their actual information preferences and the content they generate, rather than using only relational information. This would help ensure that only users that are actually similar to each other (i.e. have similar perspectives on similar content) make up each community. We use LLMs and human interaction to do this validation. First, we summarize each user in the community using LLM's, mapping their profile to a short statment capturing their views. This representation facilitates the human validation step, in which we ask humans to determine which users are similar, and by that form validated communities (experiments show this is hard for LLMs, but simple for humans). We then expand the validated communities, adding new users to them, based on model clustering. These assignments are again validated based on users' summaries, but this time using LLMs to compare new users to users in human validated communities (we use human interactions as training examples for few-shot user similarity detection, a simpler LLM task). For users not assigned to validated communities, the above process is repeated, expanding the # of communities. Fig. 1 shows an overview. In short, humans interact to help form initial communities, which are then automatically expanded using graph and LLM similarity.

In summary, we make the following contributions: **(1)**: We formulate the task of interactive news media profiling, by presenting a framework to build information communities. **(2)** We take advantage of graph model, LLM, and human knowledge to perform this interactive task rapidly. **(3)** We evaluate on one of the most challenging news media profiling settings, emerging news events, showing how our interactive framework leads to performance improvements with less than 5 human interactions. More generally, our framework can be used to rapidly profile news media, without any additional labeling and minimal human effort.

Sec. 3 describes our graph, Sec. 4 our interactive framework, Sec. 5 results, and Sec. 6 analysis.

2 Related Work

Over the last few years, there has been a large interest in profiling news media. Baly et al. and Sakketou et al. proposed datasets for political bias detection, while (Li and Goldwasser, 2019; Liu et al., 2021, 2022) study methods for it. Further, fake news detection has also been a hot research topic, studied in graphs (Nguyen et al., 2020; Mehta et al., 2022; Yang et al., 2023), cross-domain (Huang et al., 2021; Zhu et al., 2022a,b; Mosallanezhad et al., 2022), and low-resource (Lin et al., 2022) settings, amongst others.

We focus specifically on news media profiling in the emerging events setting, which is extremely challenging, as test data is unseen and does not interact with any train data. Thus, this setting is ideal for us to measure the benefits of human interactions. It also has received a lot of recent attention due to its' challenge (Liu and Wu, 2018; Li et al., 2022), and we hypothesize that some of these methods can be combined with our work.

Large Language Models (LLMs) have been applied to many tasks, as they can capture lots of knowledge (Qin et al., 2023). However, LLMs still cannot reason well, and thus struggle on harder tasks like fake news detection (Whitehouse et al., 2022). We instead use LLM's successful properties to amplify the impact of human interactions.

Interactive ML has also been studied and applied to many tasks (Wu et al., 2022; Dalvi et al., 2022; Kwon et al., 2023; Ramamurthy et al., 2023; Pacheco et al., 2022, 2023; Mehta and Goldwasser, 2023). Building information communities is also a popular research area, whether it be through embeddings or DNN modeling approaches (Cavallari et al., 2017; Su et al., 2022). Prior work also shows misinformation spreads in communities (Bessi et al., 2016). We propose to interactively build information communities, by humans interacting with LLMs and graph models. We discuss more related works in App. E.

3 Graph Overview

In this paper, we focus on detecting the political bias and factuality of news media sources, which we call **news media profiling**. We model sources for factuality and political bias on a 3-point scale: *low*, *mixed*, and *high* factuality, or *left*, *center*, and *right* political bias. More details about this task setup and its importance are in (Baly et al., 2020) and Appendix A.1.

We use the public graph-based social media analysis model from Mehta et al.[1], which they trained for fake news source detection. As we also evaluate news source bias detection, we train the graph for both objectives, on Twitter data. We refer to Mehta et al. for the details of this graph model, but briefly explain it here. Sec. 4 explains our interactive protocol for identifying information communities.

The model uses a heterogeneous graph, encoded using Relational Graph Conv. Networks (R-GCN), to capture the relationships between sources, articles, and users. Based on the R-GCN representations, factuality and bias of news sources can be predicted. R-GCNs allow us to better capture relationships in the graph, such as a source represented in part by the users that follow it (which are also represented by their relationships to other nodes).

Graph Creation using Twitter Social Context: Our graph (Fig.1a, Mehta et al.) consists of 3 node types: News Sources (S), Articles (A) they publish, and Twitter Users (U) that interact with sources and articles. We connect sources to articles they publish, via edges. Further, users are connected to sources and other users they follow, and articles they propagate (retweet or tweet the link of). The Twitter users provide the social information in the graph, which we later aim to better learn.

Graph Training using Social Context: Similar to Mehta et al., we train a R-GCN (Schlichtkrull et al., 2018) to learn the graph. We train the classification objective of both fake news source detection and news source bias detection, using a separate Fully Connected (FC) layer for each, optimizing them jointly by summing the losses. Once the model is trained, we can use it to obtain meaningful node embeddings for every node in the graph, and profile news sources. More details: App. A.3.

4 Interactive Approach

While the above graph-based model proposed by Mehta et al. achieves strong performance on fake news source detection when evaluated in transductive settings (test data seen at training time), it struggles in the fully inductive settings (test data unseen), and in general performs well below human baselines. We thus propose an interactive approach, combining the strengths of graph models, large language models (LLMs i.e. GPT-3), and humans, to profile content on social media better.

Our approach hinges on the fact that **if we can better model user preferences and thus user similarity, we can better model the content they**

[1] https://github.com/hockeybro12/FakeNews_Inference_Operators

propagate. This is because, similar users are likely to have similar interests, and thus share similar content, which in turn is likely to have similar levels of factuality/bias. For example, a group of users sharing content in support of lowering taxes and decreasing regulations are more likely to be right-biased (i.e. Republican) vs. left-biased (i.e. Democrat), and thus any source content they share is also more likely to be right-biased. Thus, we hypothesize that the larger the groups of users with similar content preferences we can form, the higher our performance is likely to be. Further, if we **explicitly create new graph edges between these similar groups of users**, this information will flow to other users not part of these groups, and eventually news sources, increasing classification performance.

Fortunately, while modeling user content preferences solely through AI models like LLMs is difficult (Whitehouse et al., 2022), humans can quickly determine if two users are similar, forming an initial group. Then, we hypothesize, that LLMs can be prompted using the human insight to extend the group, by asking them if other users are similar in the same ways. Thus, in this paper, we propose an **interactive framework**, taking advantage of human and LLM strengths to better model user content preferences, and improve media profiling.

Specifically, in this Sec., we discuss the interactive approach we propose to form these groups of similar users, or *information communities*. We take advantage of the strengths of trained graph model knowledge, LLM knowledge (GPT-3), and human insight to design an iterative, interactive approach: We first use the graph models' learned user similarities to form initial candidate information communities (Sec. 4.1), which are summarized by LLM's (Sec. 4.2), and then validated by a human interactor (Sec. 4.3). These validated communities are then expanded upon to include more users in Sec. 4.4, by again using graph model knowledge and LLM's. However, this time LLM's are prompted based on the human interaction, to validate user assignments to communities. This expansion step can be done iteratively, i.e. assigning additional users to validated communities. Once enough users are not assigned to existing communities, we form an additional set of human validated communities, repeating the the above process from Sec. 4.1-Sec. 4.4, and the number of total validated communities is increased (Sec. 4.5). Finally, once enough information communities are formed, we can then learn them (Sec 4.6), updating graph model parameters.

We later show that through this iterative process, minimal human interactions can lead to significant performance improvements for news media profiling. Fig. 1 and Alg. 1 shows an overview.

Algorithm 1 *Our Interactive Framework to Find Validated Information Communities*

```
Input: $U$ (Users), $U_E$ (Graph User Embeddings), V (empty list to store validated information communities)
Output: $V$ (Validated Communities)
Iteratively find information communities
while not converged do
    $c_{1...k}$ = k-means($U_E$) K-means Cluster all Users based on Graph User Embeddings
    $c_1, c_2 = max_2(\text{purity}(c_{1...k}))$ Choose the highest purity clusters, that discuss the same entity
    $s_1, s_2$ = GPT-3 summarize($c_1, c_2$) Use GPT-3 to summarize the users in each cluster
    $V$.append([human($s_1$), human($s_2$)]) Validate clusters using humans to form validated communities
    Now, iteratively expand each validated community
    while not converged do
        $c_{1...k}$ = k-means($U_E$) Again Cluster all Users based on Graph User Embeddings
        $c_{1...k}$ = KNN($c_{1...k}, V$) For each cluster, find the $m$ nearest neighbors to each validated community, that's our new cluster
        $V$ = (GPT-3($c_{1...k}, V$)) For each cluster, ask GPT-3 to assign or not assign users to validated clusters, expanding them
    end while
end while
return $V$ (Validated Information Communities)
```

4.1 Initial Communities from Graph Model

The first step in our process of forming information communities of similar users is forming candidate ones. For this, we use learned graph model knowledge, and k-means cluster all graph user node embeddings, as similar nodes will be part of the same cluster (and thus community). We keep the two (determined empirically using the dev. set) highest purity clusters, as the model is likely most confident about them, since it predicts similar users as having the same labels. To compute purity, each cluster is assigned to a class based on the most predicted user label in that cluster, and then the accuracy of this is measured. To get predicted user labels, we assign each user the label of the most common source they follow + article they tweet.

Since these communities are formed using graph learned relationships, they are likely imperfect, and should be analyzed to form better communities. Thus, we ask humans to analyze them. However, as the communities have a lot of users which would require a lot of interactions, we narrow them down. We only keep users that discuss the most common entity mentioned in the community, as discussion around this entity can represent the community's perspective. To do this, we run an Entity Recognition system (Akbik et al., 2019) on the articles each community user tweets, keeping users if they tweet an article containing the most frequent entity in the community. We now have initial model predicted information communities of users that discuss the same entity, and thus likely the same event.

4.2 Characterizing Users Using GPT-3

Before asking humans to validate communities based on user similarity, we form a textual representation for each user, that can be analyzed. While not essential, this representation captures relevant content and user preferences, making human interaction easier. To form it, we use LLMs (GPT-3), prompting them to create user summaries, as they have historically done well on this task (Qin et al., 2023). The summary for each user is formed based on their Twitter profile and a sample of their tweets related to the entity. An ex. of the prompt we designed is shown below in Tab. 1 and Fig. 2.

Format	Language
Question	What is the user discussing and what is their perspective?
Text	Bio: ... Tweet 1: ... Tweet 2:... Summary:
Output	The user is discussing...

Table 1: The question, text, and output format expected from GPT-3 in the prompt to create user summaries.

4.3 Human Interaction to Form Community

Based on the GPT-3 summaries of each user in the communities, we ask a human interactor to tighten each community, and only keep similar users. For this, **humans read the summaries, analyzing user perspective towards the entity**. We say users have the same perspective if they discuss the same entity in a similar way (i.e. all are against BLM protests).

To make this analysis easier for humans, we also provide humans with an LLM's opinion on which users are similar. While it is likely incorrect, as LLMs can't reason well about user similarity on unseen topics (see Sec. 6.2), it can help humans make their decision quicker. To get it, we feed all user summaries to a dialogue LLM, Chat-GPT[2], asking it: *Which users have the same perspective?* We use Chat-GPT instead of GPT-3 as it is better suited to respond without being prompted, and it is hard to create a general enough prompt for this.

Chat-GPT responds with a list of users that it thinks have the same perspectives, for ex.: *User a, b, c, d discuss ... while e, f discuss ...*, which the human then reads (along with the summaries) and uses to form a human validated community. For ex., the human can decide users (a, b, c) are in the same community, where d was thought to be part of it by Chat-GPT, but not by the human. An ex. of the exact text humans read is in Fig. 3.

4.4 Automatically Expanding Communities

So far, we have formed small, human validated information communities, each of a single perspective. Now, we amplify this human interaction by expanding these communities, while maintaining the same perspective. We do this by identifying other users that have the same perspective and adding them to the community. These larger communities can then be used to to profile news media better.

We first connect the users in each validated community to each other, which changes their and other users' graph embeddings, without any training. We then k-means cluster all user embeddings, ignoring any users already considered (to avoid redundancy). This forms k new, unique clusters, based on learned graph model knowledge. We hypothesize that if we can accurately map some users from each of these different clusters to validated communities, we would have a lot more information about each of those clusters, which could help news media profiling. (Aside: an alternative way to expand communities, which we do not pursue, is by assigning users that have similar embeddings as validated communities to them. However, this would just reinforce existing model predictions, as the model already believes these users similar, and thus likely not lead to better news media profiling.)

To map users from clusters to validated communities, for each cluster, we keep the top m users that have similar embeddings to the centroids of each validated community, as these are the most likely users to belong to the community. These m users are now the candidate users for expansion into the validated community.

As this user to community assignment is based only on graph knowledge, it may be imperfect, so we hypothesize to use LLMs to clean it up. While LLMs cannot reason about community assignments on unseen news events (we experimentally show

[2] https://chat.openai.com/chat

this), which is why we used humans in Sec. 4.3, we hypothesize that if prompted appropriately, they can compare user summaries on a topic, which the human communities are already centered around. Building on this, we few-shot prompt LLMs, asking them to identify user similarity and determine which of the m new users should be part of the validated community. To do this well, we prompt the LLM using a training example, created automatically from the human validated community. In it, community assignments humans chose when interacting are positive examples, while ones humans rejected are negative. LLMs now just have to make similar assignments as humans (i.e. determine if the new users are more similar to human accepted or rejected ones), a much simpler task. Tab. 2 shows an ex. of the test prompt for the community with users (a, b), where user a is assigned to the community and b is not. The same prompt format is used for the training example, except the summaries and assignments are provided based on the human interaction. Fig. 4 shows the full prompt.

Format	Language
Question	Which users have the same perspective?
Text	User A Summary: ... User B Summary: ... Related Users;;;;Not Related Users:
Output	User A;;;;User B

Table 2: The question, text, and output format expected from GPT-3 in the prompt to determine if users belong to a given information community. From the output, User A belongs, and User B does not.

4.5 Iterative Community Expansion

The above process in Sec. 4.4 of expanding the validated communities can then be repeated until all users are assigned to or rejected for validated communities, defining convergence.

In addition, we also use a subset of rejected users to form a new set of human validated information communities. To do this, we repeat the above process from Sec. 4.1-Sec. 4.4: cluster rejected users, summarize them, ask humans to form a new validated community, and then expand the validated community. After each iteration, we have an additional pair of human validated communities.

4.6 Unsupervised Graph Training

Above, when we form communities, we create new graph edges connecting users in the same community. We now further learn these edges/user relationships, by fine-tuning the graph model from Sec. 3, all without using any additional gold labeled data. For this, we train graph link prediction, which captures this new edge knowledge directly, encouraging connected nodes to have similar embeddings. We do it only on the sub-graph of content that was interacted on: the users and the articles/sources they are directly connected to. Specifically, we train connected nodes to be closer together in the embedding space, while user nodes in different communities should be farther apart.

After this training, the graph model captures the knowledge from the user communities identified by our framework, and can thus be directly used to better classify news sources for profiling. This is because, in the updated graph model, the new user embeddings directly affect the sources, through either direct or indirect edge connections.

4.7 Framework Recap

In short, we aim to build user information communities, used by the graph model for better news media profiling. Our framework first uses graph models to build candidate communities, which are validated by humans. Identifying the communities is hard for LLMs, but simple for humans and can be done in a few minutes. Then, the communities are expanded. The graph model generates candidates, which the LLM can validate, as it has training examples from the human validation, and just has to identify the same user similarity, a much simpler task. The entire process can be done iteratively and rapidly (under 10 minutes for 5 interaction steps).

5 Experiments

5.1 Evaluation Settings

We evaluate our framework's ability to improve fake news and news source bias detection. We focus on one of the most challenging settings for a graph framework, the fully inductive setting. Here, in addition to test data not being seen at training time, **all test nodes are not connected in any way to training set nodes**. For ex., users interacting with test set articles do not interact with any sources/articles/users seen at train time. While this setting is particularly difficult, as social media information learned at training time can't be directly used to improve test performance, it can occur, such as when a new bot farm spreads content.

In addition to the inductive setting, we also focus our evaluation on emerging news events, where all test data is from a specific event collected from a

Model	Baly Acc.	Baly F1	Test Acc	Test F1
Baly	71.52	67.25	-	-
Mehta R-GCN	68.90	63.72	-	-
Mehta BEST	72.55	66.89	-	-
BL: Mehta R-GCN	65.82	53.19	41.89	28.48

Table 3: Fake News Source Detection baseline Results on Baly (Baly et al., 2020) and the inductive future Black Lives Matter event (Test). Results show that despite achieving high performance on (Baly et al., 2020), the Baseline from Mehta et al. (BL: Mehta R-GCN) struggles in the inductive, emerging news events setting. This baseline is comparable to the state of the art for fake news source detection from Mehta et al. (Mehta BEST) on (Baly et al., 2020).

time period after the training time period. Not only is this one of the most common real-world applications for fake news source and bias detection, but it is also very challenging, as test data focuses on sub-events not seen at training. In this work, we evaluate two important news events: *Black Lives Matter (BLM)* and *Abortion/Feminism*.

5.2 Data Collection and Usage

Fake News and Bias Source Detection: In order to evaluate our framework's ability to improve fake news and news source bias detection, we used the Media Bias/Fact Check dataset, originally collected by Baly et al.. As we focus on specific events, many of which have occurred since the dataset was originally collected, we expand it by scraping additional news sources from Media Bias/Fact Check[3]. Additionally, we scraped the data used to construct the graph in Sec. 3 (articles sources publish, Twitter users, Twitter interactions, etc.) following the process in Mehta et al.. As done in Baly et al., we label news sources on a 3-point factuality and 3-point bias scale: *high*, *mixed*, or *low* factuality and *left*, *center*, or *right* bias. Dataset details, including statistics for number of sources is in App. D and Tab. 6. Our code and anonymized data is available.[4]

Events: For each event that we tested on (Black Lives Matter and Abortion/Feminism), we scraped data for 2 different time periods (01/02/2019 - 06/01/19; 06/02/19 - 05/06/22), searching relevant hashtags on Twitter. These time periods also cover a broad range of sub-events, allowing us to test how our models would do on emerging news events. To learn the graph model for fake news and bias source detection from Sec. 3, we used the first period and a subset of data from Baly et al. (training the model on the event and general news). The other time period is our test data, and forms a fully inductive graph, where none of the nodes in the test graph are connected to training set nodes, making it hard.

5.3 Evaluation

We evaluate our models primarily on Accuracy and Macro F1 score (the dataset is unbalanced), for sources. We also evaluate the # of users and sources interacted on, the total # of edges added by all interactions, the # of expansion rounds done (defined in Sec 4.4), and the # of interactions done.

5.4 Baselines

Our first baseline is the strong graph based fake news source detection model from Mehta et al., which we also trained for and evaluated on bias detection. They also compared to multiple baselines in their work. Tab. 3 shows the performance of this model on Baly et al., but when evaluated on BLM in the inductive setting, it struggles (Tab. 3, Tab. 4).

Our second baseline is our information community detection approach without humans and LLMs, creating the communities based only on graph model embeddings (Graph Only). We k-means cluster user embeddings, and choose high purity clusters, keeping the top m similar users. We choose k=35 based on validation set performance.

Our final baseline, LLM Only, is our framework without humans, but still using LLM + graph knowledge. To remove the interaction step from Sec. 4.3. where humans form validated communities by reading summaries and Chat-GPT's assignments, we instead trust the Chat-GPT assignments and use these as the "validated" communities.

5.5 Interactive Framework Results

Results for Black Lives Matter fake news and bias source detection are in Tab. 4. Abortion/Feminism results are in Tab. 7 + Tab. 8. Results show how our interactive framework (LLM + Humans) enables minimal human interactions (details about interaction process in App. C), sometimes only one, to lead to performance improvements for these tasks, even on emerging news events without additional labels. We experimented with a varying # of validated communities and expansion rounds (using the dev. set to find the #), and all showed improvements in either Acc. or F1 score over baselines. Specifically, we see ∼33% improvement on fake news source macro F1, and ∼40% improvement on bias news source macro F1. Our best models

[3] https://mediabiasfactcheck.com

[4] https://github.com/hockeybro12/Interactive_News_Media_Profiling

Model	FN Acc	FN F1	Bias Acc	Bias F1	# Users; # Sources	# Edges	# Inter- actions
Baseline: (Mehta et al., 2022)	41.89	28.48	46.79	27.43	-	-	-
Graph Only: High Purity 2 Communities (Comms.)	43.01	28.85	46.15	28.59	25; 25	1,200	-
Graph Only: High Purity 4 Communities (Comms.)	41.89	27.23	48.71	21.83	-	-	-
LLM Only: 2 Comms, 2 Expansion Rounds	42.70	28.05	45.01	27.84	38; 63	494	-
LLM Only: 4 Comms, 2 Expansion Rounds	42.70	28.62	39.50	33.22	69; 56	1,791	-
LLM Only: 6 Comms, 2 Expansion Rounds	40.54	26.88	37.03	29.22	73; 63	1,612	-
LLM + Humans: 2 Comms, 2 Expansion Rounds	**52.51**	**38.03**	44.23	33.40	25; 26	367	1
LLM + Humans: 2 Comms, 4 Expansion Rounds	46.36	35.03	**49.35**	**45.13**	72; 56	1,087	1
LLM + Humans: 4 Comms, 2 Expansion Rounds	43.01	32.36	47.43	32.00	55; 43	808	2
LLM + Humans: 6 Comms, 2 Expansion Rounds	41.34	32.36	48.07	33.91	82; 61	1,696	3

Table 4: Fake News (FN) and Bias Source Detection on Black Lives Matter: We evaluate Test Set Accuracy, Macro F1, the # of users and sources directly connected, the # of edges created, and the # of human interactions are performed (each forms 2 validated communities). Results show that our proposed approach, the human interaction models (LLM + Humans, last group), achieve improvements over all other models in Acc. and/or F1. Also, human interactions are critical, as LLM Only models (third group; they still use graph insight) do not achieve significant improvements over baselines (first and second group). Moreover, our best performance is with only 1 single human interaction, creating 2 communities and then expanding them (2 expansion rounds for fake news source detection and 4 for bias source detection).

for each task and each event only needed up to two human interactions, showing the benefit of our framework to amplify human interactions. Also, all human interaction models outperform all non-human baselines, including LLM Only, showing that both LLM and human insight (to sort out LLM inconsistencies) is critical for news media profiling.

In summary, these results shows how we are able to successfully decompose the task of finding information communities: taking advantage of graph, LLM, and human strengths, to successfully profile news media, even on emerging news events.

6 Discussion

In this section, we evaluate our Black Lives Matter interactive framework (Sec. 5.5) learned information communities. We begin by analyzing the cohesiveness of the communities, first human (Sec.. 6.1) and second automatically (App. F.1). We then show why human interactions are critical (Sec. 6.2). Finally, we analyze the communities themselves, analyzing the topics discussed (App. F.2).

6.1 Human Interactor Analysis

In this section, we manually analyze our human interaction process, by asking the interactor how many candidate users for each information community they used and did not use to represent it, in each human community validation round. Results in Tab. 5 show that as more interaction steps occur, the candidate users become more similar, as humans reject less users. This shows how our interactive process improves the model's understanding of the social media framework.

Interaction Round	Users Accepted	Users Rejected
1	3	9
2	3	6
3	4	2
4	6	1

Table 5: The number of users accepted and rejected by human interactors in each new community creation step. As more interactions occur, the # of rejected users decreases, as the graph model learns to better capture similarity. Note that the # of users presented to humans changes based on cluster sizes.

6.2 Importance of Humans

As the results in Sec. 5.5 show, the human interaction step is critical to improve news media profiling performance. This is because LLM's (i.e. Chat-GPT) cannot accurately capture user similarity, particularly for new news events, which leads to non-cohesive communities. However, humans with general world knowledge can easily determine this. As an ex. in Fig. 3 (more in App. F.3), Chat-GPT responds vaguely that all users share the same perspective, when User 3 is clearly more hostile.

Without cohesive initial communities, the training examples used to prompt the generation of further communities in Sec. 4.4 will also be non-cohesive and thus incorrect, leading to non-cohesive expanded communities. Thus, the graph model wouldn't gain any insight about user perspectives through the communities, which is why downstream performance doesn't improve.

7 Conclusion

In this paper, we proposed a framework for interactive news media source profiling. Our framework combines the strengths of graph based news media profiling models, LLMs, and humans, to build stronger information communities. We show how without any additional labeled data, and less than 5 human interactions which can be done in under 10 minutes, we can better detect fake news and bias sources, on two separate news events, even in the most challenging setting of emerging news.

Our future work is building larger and better communities, and having more human interaction rounds. We hypothesize that more data (users, sources, articles, and human interactions) could lead to lead to better communities, as our approach can capture more perspectives on social media. This would also likely lead to a better trend in the results, leading to more consistent performance improvements as the number of interaction rounds are increased.

8 Acknowledgements

We thank the anonymous reviewers of this paper for their vital feedback. The project was partially funded by NSF award IIS-2135573 and NSF CAREER award IIS-2048001.

9 Ethics Statement

For the ethics statement, we first discuss limitations of our model (9.1), and then in Sec. 9.2 we discuss ethics for deploying our models..

9.1 Limitations

In this paper, we focus on news media profiling (fake news and bias source detection) on English and Twitter, specifically in the Black Lives Matter domain. The experimental results we presented in this paper showed our framework works in these domains/tasks. We are hopeful and believe that our framework would generalize to other domains, tasks, and topics, but we leave the investigation of this to future works.

In this paper, we also primarily focused on the evaluation setting of early detection of fake/biased news sources, where we evaluate on unseen test data that is not connected to any training set data in the graph. We believe that this is one of the most challenging settings for news media profiling, as shown by prior work. We thus believe that our framework would generalize to other news media profiling settings, including ones that are not in the early detection space. Our future work involves testing this hypothesis, by combining our frameworks with other works in the early detection space.

Our framework utilizes Large Language Models, specifically GPT-3, which the details are not yet fully known publicly. Although these models have been shown to achieve strong performance in numerous NLP benchmarks (Qin et al., 2023), we believe the community should still be careful in deploying them.

Our framework also utilizes human interactions, which in our paper are extremely simple, as humans must just read short summaries to determine similarity. Further, our framework needs an extremely small amount of human interactions. However, we still caution that in a real world deployment of our framework, we should be careful of human interactors and make sure they do not have a malicious intent and are well educated for this task. Moreover, it would be better if numerous humans provided judgements on a single interaction sample, to confirm all the interactions across multiple experts.

For our experiments, we used a single GeForce GTX 1080 NVIDIA GPU, with 12 GB of memory. As our models are largely textual based, they do not require much GPU usage, but this could change in real world settings, where lots more data is available, which could be a potential limitation. Our hyper-parameter search, mentioned was done manually, based on dev set performance. The appendix provides more model details.

9.2 Ethics

To the best of our knowledge, we did not violate any code of ethics throughout the experiments done in this paper. We reported technical details necessary to reproduce our results, and will release the code and dataset we collected, upon publication. We evaluated our model on the datasets that we collected in this paper, and was collected by prior work, but it is possible that results may differ on other datasets. However, we believe our methodology is solid and applies to any social media news profiling setting, as shown by our performance on emerging news events.

Due to lack of space, we placed some of the technical details in the Appendix section. The results we reported support our claims in this paper and we believe that they are reproducible. Any qualitative result we report is an outcome from a machine learning model that does not represent the authors' personal views.

In our future dataset release, we include sources, users, and articles, so that our experiments can be replicated. Each are in English, and are public information. We map each to an ID, for anonymity, and release Article textual representations. Article texts are available for academic use, and can be provided by requesting the authors and agreeing to appropriate conditions.

Our framework in general is intended to be used to profile news media sources, and help identify the spread of misleading or perspective changing content on social media. While our framework could be used to build better methods of avoiding fake news/bias detection by ML systems, our interactive framework can guard against that as well.

In general, we caution that our models and methods be considered and used carefully, as in an area like news media profiling there are great consequences of wrong model decisions, such as unfair censorship and other social related issues. Further, it is possible our models are biased, and this should also be taken into consideration. An important future work is to investigate our models, interpreting them and understanding their predictions even better than the analysis showed in the Discussion section of this paper.

The interactive setting we proposed was successful in this paper, particularly because the interactions were simple. However, in the real world, there could be biased interactors with malicious motives, and that is an important thing to consider when dealing with fake/bias news source detection systems.

These and many other related issues are things to consider when using models such as the ones proposed in this work.

References

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 *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, Brussels, Belgium.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20.

Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics*, 225:2047–2059.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sandro Cavallari, Vincent W Zheng, Hongyun Cai, Kevin Chen-Chuan Chang, and Erik Cambria. 2017. Learning community embedding with community detection and node embedding on graphs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 377–386.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9465–9480.

Yosh Halberstam and Brian Knight. 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of public economics*, 143:73–88.

Yinqiu Huang, Min Gao, Jia Wang, and Kai Shu. 2021. Dafd: Domain adaptation framework for fake news detection. In *International Conference on Neural Information Processing*, pages 305–316. Springer.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward design with language models. In *The Eleventh International Conference on Learning Representations*.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks forpolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604.

Ke Li, Bin Guo, Siyuan Ren, and Zhiwen Yu. 2022. Adadebunk: An efficient and reliable deep state space model for adaptive fake news early detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1156–1165.

Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2543–2556.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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*.

Nikhil Mehta and Dan Goldwasser. 2023. Interactively learning social media representations improves news source factuality detection. *arXiv preprint arXiv:2309.14966*.

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.

Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, pages 3632–3640.

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.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

María Leonor Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. 2022. A holistic framework for analyzing the covid-19 vaccine debate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5821–5839.

María Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. 2023. Interactive concept learning for uncovering latent themes in large text collections. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5059–5080.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations*.

Bhavtosh Rath, Xavier Morales, and Jaideep Srivastava. 2021. Scarlet: explainable attention based graph neural network for fake news spreader prediction. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 714–727. Springer.

Bhavtosh Rath, Aadesh Salecha, and Jaideep Srivastava. 2020. Detecting fake news spreaders in social networks using inductive representation learning. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 182–189. IEEE.

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), pages 3982–3992.

Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri Jacques Geiss, Paolo Rosso, and Lucie Flek. 2022a. Factoid: A new dataset for identifying misinformation spreaders and political bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3231–3241.

Flora Sakketou, Joan Plepi, Henri-Jacques Geiss, and Lucie Flek. 2022b. Temporal graph analysis of misinformation spreaders in social media. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 89–104.

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.

Xing Su, Shan Xue, Fanzhen Liu, Jia Wu, Jian Yang, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Di Jin, et al. 2022. A comprehensive survey on community detection with deep learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

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

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, et al. 2019. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chenxi Whitehouse, Tillman Weyde, Pranava Madhyastha, and Nikos Komninos. 2022. Evaluation of fake news detection with knowledge-enhanced language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1425–1429.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*.

Sin-han Yang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Entity-aware dual co-attention network for fake news detection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 106–113.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2020. Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5444–5454.

Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E Gonzalez. 2023. The wisdom of hindsight makes language models better instruction followers. In *International Conference on Machine Learning*, pages 41414–41428. PMLR.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022a. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022b. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.

A Experimental Settings

A.1 Profiling News Media Sources Task Definition and Importance

In this paper, we focused on detecting the political bias and factuality of news media sources, which we call **news media profiling**. We focused on these tasks as stopping misinformation is critical, and politically biased content can sway beliefs and affect important real-world events, such as political elections (Vosoughi et al., 2018).

We model sources for factuality on a 3-point scale: *low*, *mixed*, and *high*. Similarly, we model sources for political bias on a 3-point scale: *left*, *center*, and *right*. More details about this task setup and its importance can be found in (Baly et al., 2020). Below, we also briefly discuss the importance of profiling the source itself.

Focusing on profiling the source itself, rather than the content, can have several benefits, which is why we focused on this task work: **(1)** Most sources publish a large amount of content, and knowing the facutality/bias of the source can give us insight about all the content they publish. Moreover, any new content that the source publishes is more likely to have a similar factuality/bias as the source's historical content. Thus, knowing the source's historical level of facutality/bias can provide insight about the new content, which can help in rapidly profiling it. For example, a source publishing mostly *left* biased content in the past is more likely to be *left* biased in the future. **(2)** There are a lot more sources than content on social media, so it can be easier to accurately profile sources, and our framework makes this task even easier. As mentioned above, doing this can then provide insights on all the content the source publishes. **(3)** There are many new news sources arising daily, so manually profiling all of them is extremely difficult. Thus, developing an automated system is critical.

A.2 Graph Initial Embeddings

We followed the released code and data from Mehta et al., so we use their exact node embedding representations as our initial graph embeddings. The Twitter embedding is a 773 dimensional vector consisting of the SBERT (Reimers and Gurevych, 2019) RoBERTa (Liu et al., 2019) representation of the user profile, consisting of features such as: user bio, user verification status, number of user followers/following, how many tweets they post, and how many likes their tweets have received. We also used similar features for YouTube embeddings for each source: number of likes, dislikes, and number of comments on their videos. For the article feature vector, it was also a SBERT RoBERTa textual embedding.

A.3 Models and Training

We used the ChatGPT models available via the OpenAI API[5] as of February 2023.

For our graph models, we used the publicly released code and hyper-parameters of Mehta et al., which uses the PyTorch (Paszke et al., 2019) and DGL (Deep Graph Library) (Wang et al., 2019) libraries in Python. The R-GCN has 5 layers, 128 hidden unites, a batch size of 128, and learning rate 0.001.

Our models are trained using a 12GB Titan XP GPU card, and intial training takes 2 hours. Link prediction training after human interactions is very quick, and can complete in under 30 minutes. Further, doing the expansion step completes in under 30 minutes. It took under 10 minutes for the humans to do all the interactions.

We used the development set to evaluate model performance, and choose the best hyper-parameters for our experiments.

Our models are trained for source classification, using a separate Fully Connected (FC) layer for each fake news source classification and political bias source classification. The R-GCN (Schlichtkrull et al., 2018) model we use creates contextualized graph embeddings for each node in the graph. For example, source embeddings are affected by users and articles they are directly or indirectly connected to. This is why our approach to learn user communities, which leads to better user embeddings (i.e. users with similar perspectives are closer together), leads to better source embeddings.

After being passed through the FC layer for classification, the R-GCN source embeddings are then passed through the Softmax activation function, and finally used to predict the source label. The model is trained like in Mehta et al., using a categorical cross-entropy loss, where the gold training labels are factuality or political bias. The two source classification tasks are trained jointly, summing their losses.

[5] https://openai.com/blog/openai-api

B Prompts for GPT-3

In this section, we describe the prompts we use for GPT-3 and Chat-GPT, to utilize the large amounts of pre-trained knowledge contained in these models to help us in our interactive framework. It has been shown (Brown et al., 2020; Wei et al., 2022) that prompting can help utilize LLMs like GPT-3 for many NLP tasks, as they cannot be trained directly. While these models cannot do complex reasoning to determine if sources are fake or biased news, they can solve simpler tasks, such as determining if users have the same perspectives, which can help with our tasks, that we built into our interactive framework. Specifically, they can summarize users based on their content (Fig.2), help humans analyze perspectives (Fig. 3), and determine community membership based on similarity (Fig. 4),

All the prompts used in this paper are human designed.

C Human Interaction Details

We used a single human interactor for all our experiments, who was a computer science PhD student, of Asian American descent. The student was compensated in research credit hours, as the interaction process was part of their research credit work.

We used only a single human due to the fact that our interaction process is very simple, as humans only read a few user summaries and determine content similarity, which in all cases we evaluated on is fairly straight-forward. Moreover, the amount of interactions done in this work was very small, as only 5 interactions could lead to the formation of 8 information communities and connect different communities together. The human we used was also expertly trained. Thus, we did not explore using additional interactors in this work, and it is something we leave for future work. For example, other setups could use multiple experts for human interactions, taking their majority vote as the final interaction.

D Dataset Statistics

Table 6 shows the statistics of the number of sources for the Black Lives Matter and Abortion/Feminism event we evaluate on.

To collect the data for Black Lives Matter and Abortion/Feminism, we searched hashtages on Twitter. The hashtags/search terms we used for the *Black Lives Matter* event were: *Black Lives Matter, BLM, blacklivesmatter, Floyd, George Floyd.* The hashtags/search terms we used for the *Abortion/Feminism* event were: *abortion, feminism, womenempowerment, womensrights, metoo, prolife, prochoice*.

E Related Work Cont.

We now discuss additional related works we didn't cover in the main paper, due to space. Several additional works aim to analyze fake news spreaders on social media, some of them using graphs (Rath et al., 2020, 2021; Sakketou et al., 2022b).

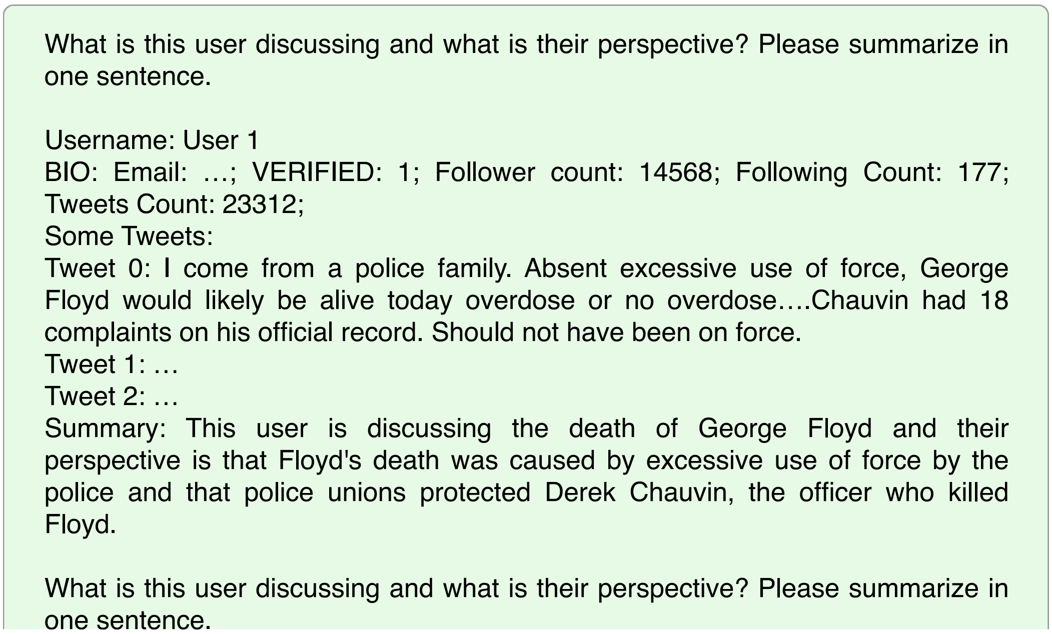
E.1 Impact of Information Communities for News Media Profiling

Prior work has shown that misinformation tends to spread in groups on social media (Bessi et al., 2016; Halberstam and Knight, 2016; Cinelli et al., 2021). Specifically, they show that like-minded users tend to form groups, that biased/false information reaches the users in these groups more quickly, and that these groups are more likely to be biased/spread misinformation.

This motivates our ideas to build better information communities for fake news and bias source detection. We hypothesize that if we can identify these like-minded users, i.e. our information communities, then we can more easily identify their users' bias and likeliness to spread misinformation. This knowledge can then be used to profile news media sources better, especially through our graph model, where users, articles, and sources are directly connected and vary in similarity to each other. Thus, we aimed to build better information communities, and took advantage of minimal human interactions to do this.

E.2 Humans Interacting with LLMs

Humans interacting with LLMs has gained popularity recently. One direction is Reinforcement Learning for Human Feedback (RLHF) (Bai et al., 2022; OpenAI, 2023), where humans preferences are used to train a Reinforcement Learning Agent reward model, which can then be used to improve the LLM. An extension of this, which needs significantly less human interactions, is having humans provide a few training examples to a LLM prompt, which can be used by the LLM to generate rewards (Kwon et al., 2023). In another direction, human interactions can be used to generate LLM prompt instructions to better solve a variety of NLP Tasks (Zhang et al., 2023).

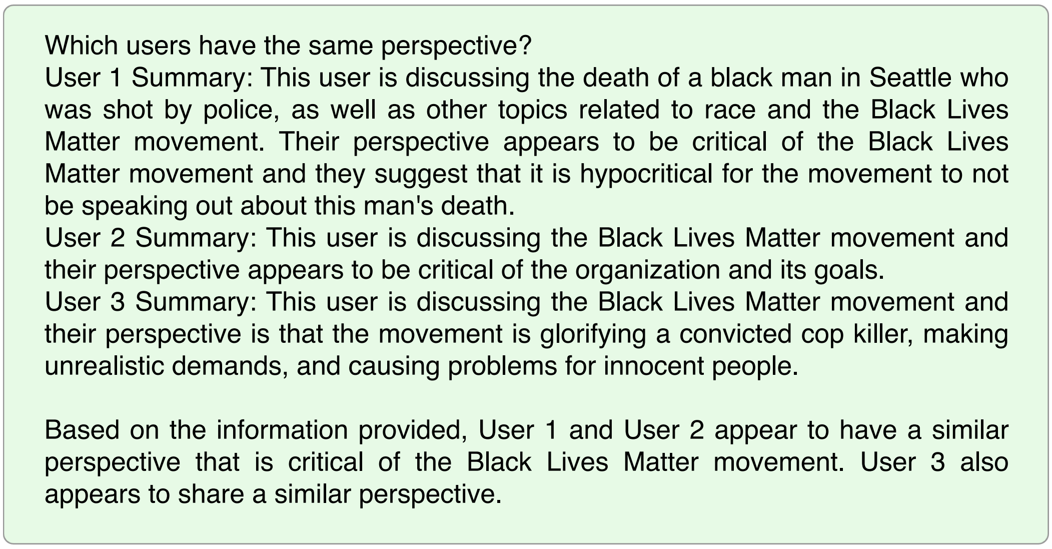


What is this user discussing and what is their perspective? Please summarize in one sentence.

Username: User 1
BIO: Email: …; VERIFIED: 1; Follower count: 14568; Following Count: 177; Tweets Count: 23312;
Some Tweets:
Tweet 0: I come from a police family. Absent excessive use of force, George Floyd would likely be alive today overdose or no overdose….Chauvin had 18 complaints on his official record. Should not have been on force.
Tweet 1: …
Tweet 2: …
Summary: This user is discussing the death of George Floyd and their perspective is that Floyd's death was caused by excessive use of force by the police and that police unions protected Derek Chauvin, the officer who killed Floyd.

What is this user discussing and what is their perspective? Please summarize in one sentence.

Figure 2: An example of the prompt we used to determine the user summary. Based on their bio, meta-data, and tweets, we create a summary.



Which users have the same perspective?
User 1 Summary: This user is discussing the death of a black man in Seattle who was shot by police, as well as other topics related to race and the Black Lives Matter movement. Their perspective appears to be critical of the Black Lives Matter movement and they suggest that it is hypocritical for the movement to not be speaking out about this man's death.
User 2 Summary: This user is discussing the Black Lives Matter movement and their perspective appears to be critical of the organization and its goals.
User 3 Summary: This user is discussing the Black Lives Matter movement and their perspective is that the movement is glorifying a convicted cop killer, making unrealistic demands, and causing problems for innocent people.

Based on the information provided, User 1 and User 2 appear to have a similar perspective that is critical of the Black Lives Matter movement. User 3 also appears to share a similar perspective.

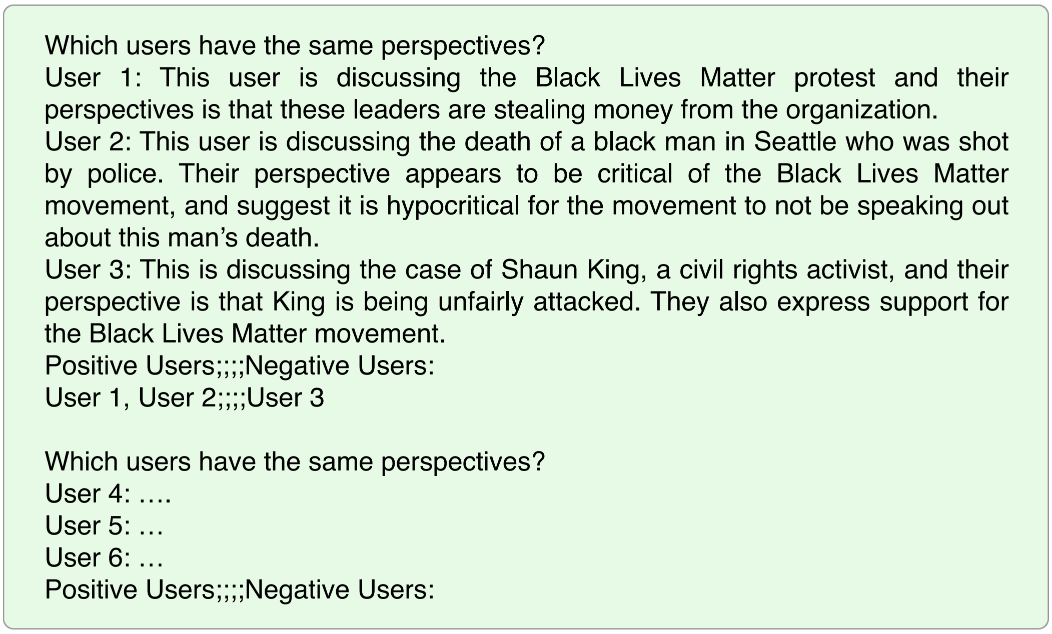
Figure 3: An example of the output shown by Chat-GPT when provided user summaries and asked to predict similarity. Note how often times the output can be vague, which is why human interactions are necessary.

In contrast, in this work, we had humans interact with LLM and graph knowledge to build better information communities for news media profiling. To do this, we first used LLMs to generate user and community summaries, which we presented to humans. Humans were then asked to use their judgement and reasoning skills, something that is simple for them but hard for LLMs, to form initial validated communities. These validated communities consisted of users who had similar perspectives on similar topics. Then, using graph knowledge, we generated additional candidate communities, and asked LLMs if the users in these additional communities belonged to any of the the human validated communities. As a prompt for this decision, we used the human validated community assignments.

In summary, we used LLMs to help humans (i.e. generate summaries) and amplify human interactions (i.e. determine if additional users are similar to users in any human validated communities).

F Discussion Continued

We now continue our discussion from Sec. 6, analyzing our Black Lives Matter human interaction models.



Which users have the same perspectives?
User 1: This user is discussing the Black Lives Matter protest and their perspectives is that these leaders are stealing money from the organization.
User 2: This user is discussing the death of a black man in Seattle who was shot by police. Their perspective appears to be critical of the Black Lives Matter movement, and suggest it is hypocritical for the movement to not be speaking out about this man's death.
User 3: This is discussing the case of Shaun King, a civil rights activist, and their perspective is that King is being unfairly attacked. They also express support for the Black Lives Matter movement.
Positive Users;;;;Negative Users:
User 1, User 2;;;;User 3

Which users have the same perspectives?
User 4: ….
User 5: …
User 6: …
Positive Users;;;;Negative Users:

Figure 4: An example of the prompt we used to determine community membership for one of the human validated information communities. We use the first paragraph as a 1-shot example, to prompt the model. User 1 and 2 are both critical of the Black Lives Matter movement protests, and thus part of the same community, while User 3 is in support of it, and thus shouldn't be in the community. Based on this, we prompt GPT-3 with additional users (in this case User 4, 5, and 6), and ask it to determine which users belong in the community and which do not.

Dataset	**Low Factuality or Left Bias**	**Mixed Factuality or Center Bias**	**High Factuality or Right Bias**
Black Lives Matter Bias	49	56	74
Black Lives Matter Factuality	35	45	76
Abortion + Feminism Bias	38	50	91
Abortion + Feminism Factuality	49	72	82

Table 6: Number of sources in our datasets for each emerging news event we evaluate on.

F.1 Community Cohesiveness Analysis

In this sub-section, we automatically analyze how many users in each community have the same perspectives. To do this, as an approximation, we hypothesize that the communities of users with similar perspectives likely have users with the same bias label. We use bias as an *approximation* as we have gold data for it, and users with the same political bias likely have similar perspectives (i.e. right bias users likely want to lower taxes). Tab. 9 shows that even in the final expansion round (i.e. after multiple steps of human interaction + model expansion - LLM + Humans Model) users in the communities largely have the same bias label, both when chosen by humans and automatically expanded. Thus, this approximation shows that our communities are in some ways cohesive.

On the contrary, in Tab. 9, the LLM Only model doesn't have as many users having the same labels as the LLM + Humans model, showing that without human interactions it may be harder to learn user perspectives.

F.2 Human Analysis of Community Topics

We now manually analyze the information communities learned by our best performing model on the Black Lives Matter event, by looking at the top 5 user summaries, determined by user embedding similarity to the community centroid. We observe that our communities capture meaningful perspectives. One community is against the Black Lives Matter protests, believing they are causing damage and the leaders are not condemning it. Another is in support of them, as they feel police do not treat everyone fairly. Other important sub-topics are also discussed, such as George Floyd murder, Ahmaud Arbery murder, police brutality and unions. All these are important BLM related topics.

F.3 LLM Only Failure Cases

In this section, we provide a few more examples of cases where Chat-GPT couldn't find good information communities, and thus humans were needed,

Model	Test Acc	Test F1	# Users; # Sources	# Edges	# Interactions
Baseline: (Mehta et al., 2022)	36.04	23.32	-	-	-
Graph Only: High Purity 4 Communities	34.95	19.09	22; 11	176	-
LLM + Humans: 2 Validated Comms, 2 Expansion Rounds	**37.37**	**25.24**	173; 16	280	1
LLM + Humans: 4 Validated Comms, 2 Expansion Rounds	35.43	19.64	50; 32	628	2

Table 7: Fake News Source Detection on Abortion/Feminism: We evaluate Test Set Accuracy, Macro F1, the number of users and sources directly connected/connected to, the number of edges created, and how many human interactions are performed. Results show both of our human interaction models (LLM + Humans) achieve improvements over other models (Baseline and Graph Only model). Specifically, creating two human validated communities and then expanding them over 2 expansion rounds achieves the highest fake news source detection performance. The final communities have 173 users, interact directly with 16 sources, and create 280 edges. Moreover, this performance improvement is with only 1 single human interaction.

Model	Test Acc	Test F1	# Users; # Sources	# Edges	# Interactions
Baseline: (Mehta et al., 2022)	46.92	33.21	-	-	-
Graph Only: High Purity 4 Communities	47.86	34.02	22; 11	176	-
LLM + Humans: 2 Validated Comms, 2 Expansion Rounds.	46.92	**40.33**	173; 16	280	1
LLM + Humans: 4 Validated Comms, 2 Expansion Rounds	**51.39**	38.05	50; 32	628	2

Table 8: Bias News Source Detection on Abortion/Feminism: We evaluate Test Set Accuracy, Macro F1, the number of users and sources directly connected/connected to, the number of edges created, and how many human interactions are performed. Results show both of our human interaction models (LLM + Humans) achieve improvements over other models (Baselines and Graph Only). Specifically, creating four human validated communities and then expanding them over 2 expansion rounds achieves the highest bias news source detection accuracy. The final communities have 50 users, interact directly with 32 sources, and create 628 edges. Moreover, this performance improvement is with only 2 human interactions.

Comm. #	Dominant LLM + Humans Label	LLM Only: % Of Users with Dominant Label	LLM + Humans: % Of Users with Dominant Label
1	Right	∼50%	∼60%
2	Right	∼37%	∼58%
3	Right	∼43%	100%
4	Center	40%	50%
5	Left	∼71%	∼66%

Table 9: At the final expansion round (i.e. after multiple steps of human interaction + model expansion) the majority of each community's users for the LLM + Human Interaction Model (last column) have the same gold bias label as the dominant one in the community, showing high cohesiveness (at least in gold bias label). On the contrary, the LLM Only model (third column) has a lower percentage of users with the same gold bias label, showing that without human interactions it is harder to learn user perspectives, at least based on this approximation analysis.

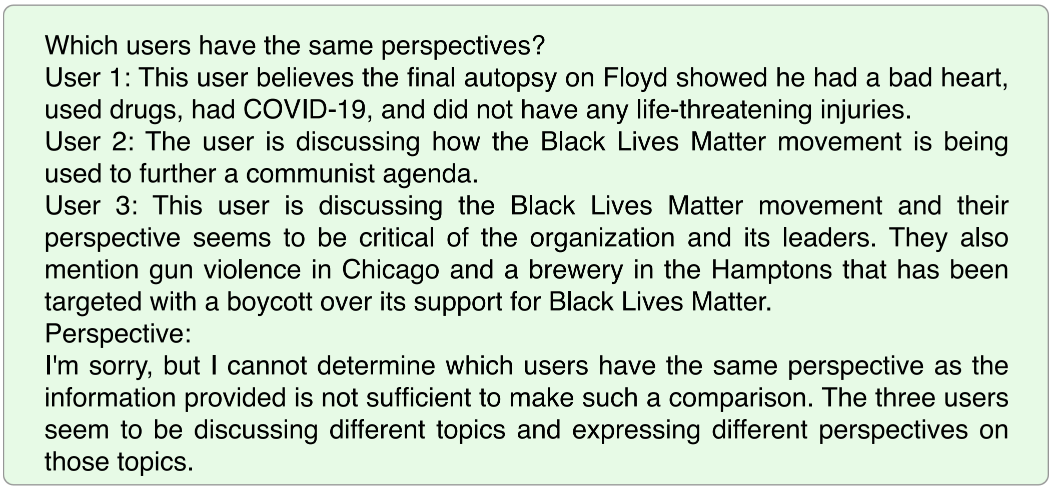
as in Sec. C. The examples are shown in Fig. 5 and Fig. 6, and the captions of the figures describe the failures.

While in this paper we experimented with GPT-3 and Chat-GPT as our LLMs of choice, we hypothesize that our results and framework would hold true for other strong LLMs as well. First, other LLMs are likely to also struggle at finding information communities on unseen data, as it is very challenging for AI models to make inferences on data and topics they have never seen before. Second, it is likely that other LLMs that perform well on determining text similarity can be used with our framework (Zhao et al., 2023). We primarily used LLMs to expand communities, by asking them to determine if new users are similar to users chosen by humans to be part of validated communities. In some ways, this is a text similarity problem, as user summaries are compared. We leave the further investigation of the choice of LLMs for use with our framework to future work.

F.4 Finer Cases

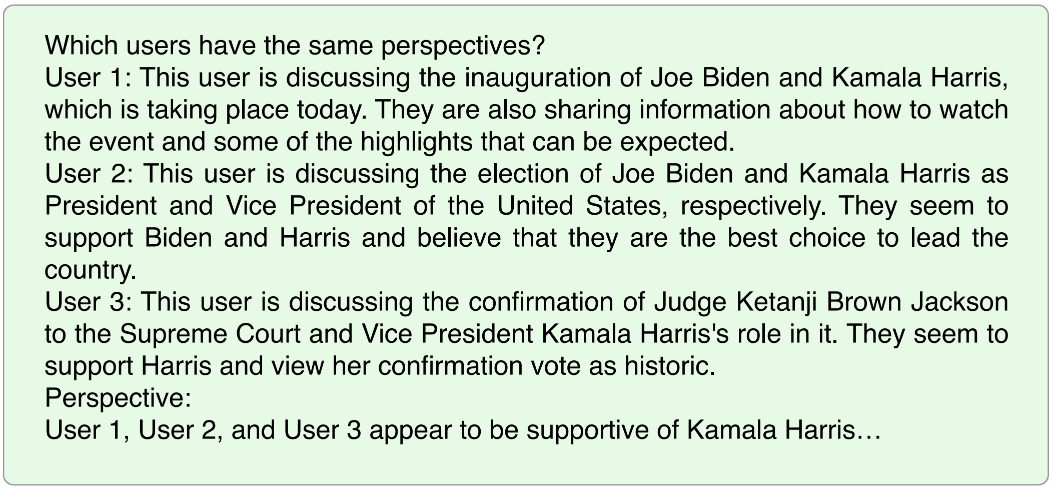
In this sub-section, we discuss some finer cases that could occur in our community analysis, and how they would or would not affect our framework.

Communities spreading both fake news and biased content: A group of users that spreads



Which users have the same perspectives?
User 1: This user believes the final autopsy on Floyd showed he had a bad heart, used drugs, had COVID-19, and did not have any life-threatening injuries.
User 2: The user is discussing how the Black Lives Matter movement is being used to further a communist agenda.
User 3: This user is discussing the Black Lives Matter movement and their perspective seems to be critical of the organization and its leaders. They also mention gun violence in Chicago and a brewery in the Hamptons that has been targeted with a boycott over its support for Black Lives Matter.
Perspective:
I'm sorry, but I cannot determine which users have the same perspective as the information provided is not sufficient to make such a comparison. The three users seem to be discussing different topics and expressing different perspectives on those topics.

Figure 5: LLM Failure Case: In this case, the LLM (Chat-GPT) can't find any communities, but it is clear that at least User 2 and User 3 should be in the same community, as they are both against the Black Lives Matter movement.



Which users have the same perspectives?
User 1: This user is discussing the inauguration of Joe Biden and Kamala Harris, which is taking place today. They are also sharing information about how to watch the event and some of the highlights that can be expected.
User 2: This user is discussing the election of Joe Biden and Kamala Harris as President and Vice President of the United States, respectively. They seem to support Biden and Harris and believe that they are the best choice to lead the country.
User 3: This user is discussing the confirmation of Judge Ketanji Brown Jackson to the Supreme Court and Vice President Kamala Harris's role in it. They seem to support Harris and view her confirmation vote as historic.
Perspective:
User 1, User 2, and User 3 appear to be supportive of Kamala Harris…

Figure 6: LLM Failure Case: In this case, the LLM (Chat-GPT) considers all 3 users as similar and in the same community. However, User 1 doesn't belong, as they are likely just reporting on the news of the inauguration. The LLM gets confused over the word "highlights", which could be positive or negative in this situation.

both fake news and biased content likely wouldn't affect our approach/results. Through the approach in Sec. 4, the group would be identified as a single information community. Then, the R-GCN graph model would learn that this community is both a fake news and political bias spreading information community.

Source Label Inconsistencies: In this paper, we obtained source labels for both factuality and political bias from Media Bias/Fact Check[6], a popular source fact-checking and source political bias detection website. This website typically holistically evaluates sources, so it is likely a majority of the content we scraped for our experiments follows the factuality/bias label provided by Media Bias/Fact-Check. However, it is possible that some sources have different levels of factuality/bias for different events that Media Bias/Fact Check doesn't capture (i.e. a source could be labeled as "mixed" factuality but be completely factual on Black Lives Matter news). While we leave the analysis of this to future work, we hypothesize that these cases are rare and thus unlikely to significantly affect our results. More importantly, it's possible that our framework can actually make the correct prediction on these incorrectly labeled sources. This can happen as our formed information communities typically span multiple sources and are very cohesive, so they are likely to be accurate. Once we form our communities, we train the graph for link prediction, so we aren't actually using the source labels but rather we are capturing user perspectives through training. Thus, if some sources are labeled incorrectly and that leads to an initially biased graph model, our interaction process can actually fix that by helping us learn a better model, improving performance on

[6] https://mediabiasfactcheck.com

those incorrectly labeled sources.

Human Interaction Inconsistencies: As mentioned in the main paper, our interaction task is extremely simple, as humans just have to determine user similarity by reading a few short spans of user text. This distinction is very clear-cut, as users that are borderline similar should not be placed in the same community. Thus, different human interactors are likely to make the same decisions, no matter their backgrounds/beliefs/etc., because identifying this level of similarity is a fairly simple task. Moreover, there are multiple ways to ensure that the interactions are done accurately, such as hiring multiple experts and taking their majority vote. However, still, in this section we analyze the situation in which humans incorrectly choose users as similar when they are not, or vice versa.

The strength of our framework is forming cohesive information communities through the interactions, where users have similar perspectives. Assuming an interaction lead to a community that wasn't cohesive, it is likely that the trained graph model would learn to ignore it, as it wouldn't gain any significant insight from this "random" community. Thus, this "incorrect" interaction is not likely to significantly hurt our model. Further, on a large scale over many interactions and lots of communities formed, a few "incorrect" interactions is unlikely to make huge negative difference to our approach, due to the fact that we always train for them using link prediction, so the model can learn to ignore it if necessary. For example, even though the link prediction training objective would pull the non-similar users in this "incorrect" community closer together, the content these users are connected to in a different "correct" community would still be pulled closer together. If there are more "correct" communities than incorrect, then even the users in the "incorrect" community would be indirectly affected by other communities and end up with the correct representations (i.e. farther apart if they are non-similar).