The PGNSC Benchmark: How Do We Predict Where Information Spreads?

Alexander K. Taylor[1†], **Wei Wang**[1]
[1]Department of Computer Science, University of California, Los Angeles
{ataylor2, weiwang}@cs.ucla.edu

Abstract

Social networks have become ideal vehicles for news dissemination because posted content is easily able to reach users beyond a news outlet's direct audience. Understanding how information is transmitted among communities of users is a critical step towards understanding the impact social networks have on real-world events. Two significant barriers in this vein of work are identifying user clusters and meaningfully characterizing these communities. Thus, we propose the PGNSC benchmark, which builds information pathways based on the audiences of influential news sources and uses their content to characterize the communities. We present methods of aggregating these news-source-centric communities and for constructing the community feature representations that are used sequentially to construct information pathway prediction pipelines. Lastly, we perform extensive experiments to demonstrate the performance of baseline pipeline constructions and to highlight the possibilities for future work. Our code and data can be found here: https://github.com/ataylor24/PGNSC.

1 Introduction

Social media platforms have become a crucial part of the information dissemination ecosystem. By allowing users to choose whom they share their so-called "content feeds" with, these platforms have created an environment in which the reach of information is amplified.

Pew Research (Forman-Katz, 2022) reported that roughly half of American adults regularly consume news through social media and that 13% prefer to get their news through social media, which increases to 33% for adults under 30. Further surveying suggests that adults under 30 place as much trust in news gathered from social media as from traditional news outlets (Liedke, 2022). Because of this, traditionally "offline" news sources now have dedicated social media accounts that seek to propagate their content to the growing demographic of adults using social media for news consumption.

Given the relevance of this problem, we seek to establish a benchmark that will establish a foundation which future work may use to develop methods to identify the audiences of influential news sources and predict the flow of information. Because social media users rarely explicitly disclose which news sources they get information from, we must identify the audience, or community, of each news source we consider. We will focus on two benefits of aggregating users into communities (Zhang et al., 2019a; Lancichinetti et al., 2008).

The *first* benefit of community aggregation is that it allows us to leverage the structure present in the data to predict the behavior of a set of closely connected users (Lancichinetti et al., 2008). Using this principle, we are able to circumvent the low-resource setting of user-level prediction caused by inconsistent user posting schedules by predicting the behavior of *communities* of users as demonstrated in prior work (Li et al., 2022; Taylor et al., 2023). The *second* benefit is that we can leverage inductive bias to improve the representation of the community; for instance, members of communities centered around a given news outlet will likely share that outlet's political leaning (Liedke, 2022) and are likely to have their views influenced by that outlet's content. This motivates the construction of community representations that incorporate as much additional information as possible (to inform community representation).

We seek to take the first step in establishing a benchmark for information pathway prediction at the community level, which we title **P**rominent **G**lobal **N**ews **S**ources for **C**ovid-19, hereafter referred to as **PGNSC**. PGNSC is a novel, human-validated dataset built using the most influential

†Corresponding author: Alexander K. Taylor {ataylor2@cs.ucla.edu}

Findings of the Association for Computational Linguistics: ACL 2024, pages 15787–15803

global news sources for COVID-19 (PGNSC). The dataset consists of instances of news articles being posted to social media and the resulting interactions between news organization communities. We also provide a general sequential framework for building pipelines to perform information pathway prediction and define baseline methods for each pipeline stage.

For the community aggregation stage, we include several methods of aggregating communities based on prior work (Taylor et al., 2023; Komorowski et al., 2018; Romero et al., 2010) and show their impact on information pathway prediction performance. To construct community feature representations, we seek to leverage the recent advances in large language models (LLMs) and their use in enhancing graph representations by using LLMs to summarize and encode of each organization's content (He et al., 2023; Chen et al., 2023b). Because news content often includes images, we also incorporate a jointly-trained image-text encoder into the set of community node feature generation pipelines (Radford et al., 2021).

The appeal of PGNSC goes beyond providing data that can be used for analysis of patterns of information propagation as well as graph representation tasks. To the best of our knowledge, this is the first work to establish benchmark for information pathway prediction using heterogeneous graphs and and to use SOTA LLMs to enhance node feature representations to improve predicitons. We believe PGNSC is well positioned to serve as a vehicle for exploring how LLMs and graph data can be used in concert to make predictions.

2 Dynamic Graph Benchmarks

The development of graph representations has recently undergone a renaissance with the development and application of graph neural networks that benefit from data richness and complexity (Huang et al., 2023; noa; Gravina and Bacciu, 2023). This evolution has underscored the critical need for robust benchmarks in both static and dynamic graph domains, divided into real-world and synthetic datasets. Our focus herein is on dynamic graphs, which are pivotal for modeling time-evolving relationships in numerous applications.

2.1 Real-world Dynamic Graph Benchmarks

High-quality, real-world datasets are considered the gold standard for benchmarking because they closely simulate the application of models in practical scenarios. However, such data can be costly to munge and often presents issues related to missing values or other quality control measures. There are many available datasets built from real-world data and described in prior works as shown in 1 (please see Appendix 8 for a full list) (Poursafaei et al.; Huang et al., 2023; Gravina and Bacciu, 2023; Horawalavithana et al., 2022). While these benchmarks encompass a broad range of spatial attributes and temporal granularity, they are overwhelmingly skewed towards homogeneous graphs with some exceptions (Poursafaei et al.; Huang et al., 2023; Gravina and Bacciu, 2023). These works, however, have not fully exploited the potential of dynamic, heterogeneous datasets nor embraced the advancements in large language models (LLMs) for multimodal representation. Our benchmark aims to fill this gap by integrating LLM-enhanced multimodal data, setting a new stage for information pathway prediction research.

2.2 Synthetic Dynamic Graph Benchmarks

To bypass the limitations of real-world data, some work has investigated synthetic datasets designed to replicate real-world data distributions (Greene et al., 2010; Ammar and Özsu, 2014; Lancichinetti et al., 2008; Rossetti, 2017). Synthetic benchmarks, while valuable for their controllability and reproducibility, often lack the unpredictability and intricate variability found in natural datasets, especially in domains like information propagation. They may not fully capture the complexities and nuances of real-world scenarios, and thus cannot provide the same value as empirical data. Our benchmark contributes a significant, large-scale real-world dataset, aiming to provide a comprehensive tool for developing more generalizable and robust GNNs.

3 PGNSC Benchmark

In this section, we discuss the construction of the PGNSC benchmark. To our knowledge, this is the first publicly available benchmark in which multimodal information cascades (structure, text, and image data) have been aggregated into community-based information pathways.

3.1 News Data Collection

Prior to selecting the news organizations whose pathways are included in the PGNSC dataset, we identified countries from which to choose news or-

Table 1: Statistics of sampled social-media-oriented dynamic-graph datasets (Hom. denotes homogeneous and Het. denotes heterogeneous). PGNSC is the dataset we introduce in this paper, and we show the statistics of both the user-level (UL) and community-level (CL) graphs.

Name	#Nodes	#Edges	Snapshot Size (N/E)	Granularity of Timestamp	Type
FB-Forum	899	7.1k	–	Unix	Homogeneous
FB-Covid19	456	54.1k	152/13.7	Daily	Homogeneous
Reddit	11k	73.5k	–	Unix	Homogeneous
Reddit Hyperlink Network	55.9k	858.5k	–	Unix	Homogeneous
Twitter-Tennis	1k	40.3k	1k/41-936	Hourly	Homogeneous
tgbn-reddit	11.8k	27.2M	–	–	Homogeneous
Expert[1]	3.5M	34M	–	–	Heterogeneous
PGNSC (UL)	11.8k	104.4M	381.2/1322.3	Unix	Heterogeneous
PGNSC (CL)	150	104.4M	12.9/36.6	Unix	Heterogeneous

[1] While we include this dataset for the sake of a comprehensive survey, to our knowledge it is not publicly available nor used by other work (with the exception of background citations).

ganizations. We identified the 15 countries with the highest aggregate number of COVID-19 cases from May 15, 2020 to April 11, 2021. We then used the Digital News Report from the Reuters Institute as our primary source in determining the most prominent online news organizations for each country, and we supplemented with additional sources when the report did not cover a given country or required further justification. A full list of the countries and news sources identified for each country is available in the appendix. We then used the aforementioned URLs to query Google News for the data associated with the article link using the open-source package **GNews**. The data retrieved contained the article title, article text content, and the URLs for images embedded in the article, as shown in Table 4.

Twitter Data	
Number of Tweets	304k
Number of Distinct Users	1.01M
Number of Articles	34k
Avg. Number of Articles per News Source	142.09
News Data	
Number of News Sources	150
Avg. Number of Articles per Country	2.7M
Avg. Number of News Sources per Country	7.89
Avg. Number of Images per Article	7.12
Avg. Number of Sentences per Article	17.17

Table 4: Relevant statistics of the Twitter and News data.

3.2 Twitter Data Collection

We utilized the COVID-19 Twitter API to retrieve data from May 15, 2020 to April 11, 2021 (331 days) using COVID-19 keywords. From this window, we identified roughly 53 million tweets containing article URLs (hereafter referred to as source tweets) from the selected news sources (see Appendix for data breakdown). We performed weighted random sampling based on the number of article links present in each day, which yielded roughly 304k source tweets. We then scraped all retweets and reply tweets (hereafter referred to as response tweets) for each *source tweet* as well as the metadata and content of each tweet (both source and reply)[0]. Each source tweet and the subsequent response tweets were used to construct a User-Level Information Cascade.

3.3 Information Pathway Mapping

We define *communities* as sets of users aggregated around the news sources defined in PGNSC (Taylor et al., 2023). Community aggregations were performed using the engagement metrics and user interactions drawn from the user-level information pathways according to methods described in detail in a later section. Each user-level information pathway was mapped to a community-level information pathway by replacing each user with the community it was assigned to (no self-edges) as shown in Figure 2.

We define a set of node and edge types that are used to construct the graph representation in Table 5, following the format of prior work (Taylor

[0]We want to note that while engagement metrics (number of favorites, retweets, and replies) were present for all source tweets, each value was defaulted to 0 for many response tweets.

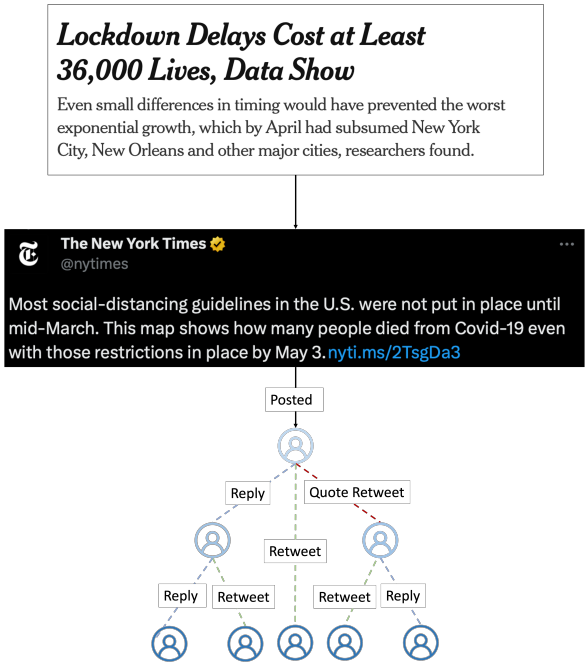

Figure 1: User-Level Information Cascade Example

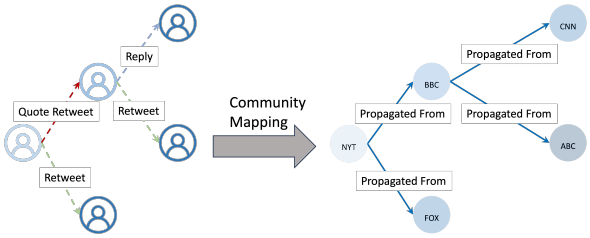

Figure 2: User-Level Information Pathway to Community-Level Information Pathway example

et al., 2023). We generalize the naming conventions for connections at the community level beyond Twitter for ease of future application to other social media platforms.

3.4 Community Aggregation Methods

In this section, we describe the aggregation methods that we used to assign users to communities centered around the aforementioned news sources. While there are many methods for unstructured community identification, we found that the majority of methods do not scale to the size of our graph for the user-level network we have constructed (Zhang et al., 2019a; Zeng and Yu, 2018; Sattar and Arifuzzaman, 2022). We have selected three heuristic-based methods. Each community aggregation method was performed using the same set of community centers and applied to the same set of user-level information cascades. The general format for the aggregation methods presented in this work is generating community scores s for each

Table 5: Definitions of Community-Level Information Pathways.

Node Types	
Information Source	An article written by a selected news source.
Community	The set of users aggregated around selected news sources.
Edge Types	
Written_by	*Information Source* → *Community*: Indicates the community that is the original author of the Information Source.
Mentioned_by	*Information Source* → *Community*: Indicates that a community authored a message containing an Information Source, starting an Information Pathway (IP).
Communicated_to	*Community* → *Community*: Represents a user from one community interacting with a message authored by a user in a different community.

user, and then assigning the user to the community for which it has the highest score, as shown in Figure 3.

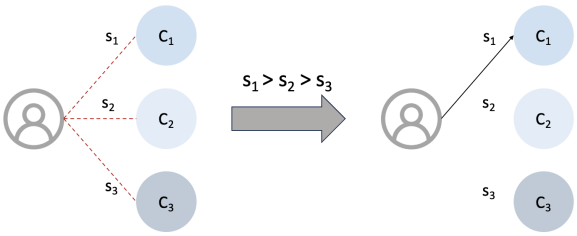

Figure 3: Illustration of community assignment. For each user, a score is generated for all communities that the user interacts with; the user is then assigned to the community for which they have the highest score.

Comm. Type	Mean	Median	Max	Min
Unbias.-Rand.	7.47k	7.48k	7.73k	7.22k
Bias.-Rand.	126.8k	80.7k	292k	17.4k
Engagement	126.7k	73.3k	328k	13.1k
Interaction	112k	77.6k	248k	31
Inf.-Pass.	18.8k	15.3k	43.3k	1.37k

Table 7: Size distribution for each community aggregation method.

3.4.1 Engagement

Following prior work on information pathway prediction (Taylor et al., 2023), we compute the Engagement score for each user with respect to each community center. This community aggregation method is designed to assign users to the commu-

nity for which their posts have the most engagement and uses collected engagement metrics: the number of favorites, retweets, replies, user IDs mentioned in the tweet, and the most recent Following/Follower ratio. Users are assigned to only the community for which they have the highest Engagement score with ties broken randomly.

3.4.2 Interaction

Following prior work in community aggregation in Twitter data (Komorowski et al., 2018), we compute the Interaction score for each user with respect to each community center (Our formulation is modified to include the number of replies). This community aggregation method is designed to assign each user to the community for which it has the highest relative importance, as determined by the impact of its interactions with other users. We first construct super-graphs of the User-Level Information Pathway Instances associated with each community center. The Interaction score is used as the edge weight for an interaction from user p and user q.

The PageRank algorithm (Page et al., 1999) is then applied to each community center's super-graph and produces a ranking of the relative importance of each user with respect to that community center. u represents the user whose rank we wish to calculate and E_u represents the set of users that have interactions with u. d is a damping factor, and $L(p)$ is the number of links from page p. Users are assigned to only the community for which they have the highest Interaction score, with ties broken randomly.

3.4.3 Influence-Passivity

The last community aggregation method that we consider follows prior work that measured the relative influence of nodes in social networks (Romero et al., 2010). This community aggregation method is designed to measure the influence that a given node exerts on its neighborhood (influence) and vice versa (passivity). For each user in a given community, we retrieve the number of URLs they posted and their one-hop neighborhood of in and outgoing interactions, which results a directed super-graph for each community. This information allows us to perform the Influence-Passivity algorithm (Romero et al., 2010) which yields a 2-dimensional influence-passivity vector, $[influence\_score, passivity\_score]$. We assign each user to the community for which they have the highest l_2-norm of the influence-passivity vector.

3.5 Random

To evaluate the effectiveness of the community aggregation methods, we include two random settings, biased-random and unbiased-random.

Biased-random. The biased-random community selection method samples from the communities whose articles a given user has interacted with. We include this setting to compare randomly selecting from a user's engagement history with using the bias described by the methods described in previous sections.

Unbiased-random. The unbiased-random community selection method randomly assigns each user to one of any community. This is the traditional random baseline and is included to evaluate the information compression properties of community assignment.

3.6 Comparison to Similar Benchmarks

We include a selected list of several dynamic heterogeneous graph benchmarks in Table 1. FB-Forum, FB-Covid19, Reddit, Reddit Hyperlink Network, Twitter-Tennis and tgbn-reddit, while rich structurally, do not incorporate node attributes beyond structural embeddings, and only FB-Covid19 and Reddit incorporate edge attributes in the form of LIWC encodings (Opsahl, 2011; Kumar et al., 2018; Panagopoulos et al., 2021; Béres et al., 2018; Huang et al., 2023). As mentioned in 1, to our knowledge, Expert is not publicly available for comparison. PGNSC is a publicly available, heterogeneous graph benchmark that incorporates LLM-enhanced multimodal graph attributes.

4 Feature Initialization Pipelines

In this section, we describe the pipelines used to generate node feature embeddings for each community. We explore the efficacy of three approaches to pipeline construction: single-stage, two-stage, and zero-shot.

4.1 Single-Stage Pipeline (Lightning)

The single-stage pipeline formulation (Lightning) illustrated in (Taylor et al., 2023) applies an encoder-only model to text that represents each community center. Lightning applies the encoder-only model, Longformer, to the concatenation of text from sampled articles to construct a representation of each community center (Beltagy et al., 2020; Taylor et al., 2023). The text we use consists

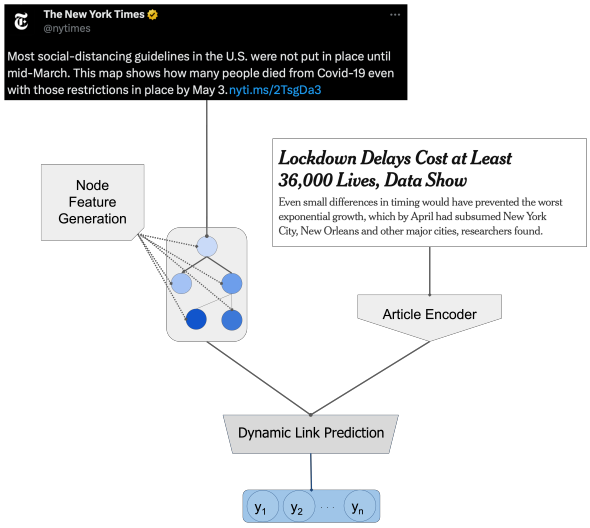

Figure 4: Illustration of the general baseline pipeline framework.

of the concatenations of article titles and article contents (illustrated in Figure 4). The embeddings generated through this process are used directly as the node feature representation of the community center to which they pertain.

4.2 Two-Stage Pipelines

Recent works have made strides in multi-document Summarization (MDS) (Xiao et al., 2022; Zhang et al., 2019b), and we seek to incorporate this into our baseline pipeline formulation. We chose to incorporate the state-of-the-art multi-document summary model, PRIMERA (Xiao et al., 2022), into the node feature embeddings generation process. This is accomplished by randomly sampling k articles from the aggregated articles available for each community to use as input to PRIMERA, which produces a text-based summary. We treat this summary as an encapsulation of the community's content and encode it using BERT to serve as the community's node features (Devlin et al., 2019; Taylor et al., 2023). We title this node feature generation pipeline PRIMERA-BERT. We also seek to explore the impacts of incorporating multi-modal data into the information pathway prediction task. Thus, we also include a pipeline that encodes the PRIMERA-generated text summary with the CLIP model (Radford et al., 2021), which is a joint text-image encoder. The process is similar to that of the PRIMERA-BERT pipeline: we randomly sample k articles *and their accompanying images* and simultaneously encode the PRIMERA-generated summary and relevant images using CLIP. We title this node feature generation pipeline PRIMERA-CLIP. We include samples of the generated summaries in the Appendix.

4.3 Zero-Shot Pipelines

Modern LLMs have made dramatic progress in the recent past and are often able to outperform fine-tuned models in low-resource settings. Thus, we also test pipelines in which we leverage the vast context of LLMs to generate zero-shot text summaries of each community using manual prompt-tuning to achieve high-quality results. The generated text summaries are encoded using both BERT and CLIP, similar to the pipelines described above. We use two of the standard LLMs, Llama-2 and GPT-4, to perform these tasks and follow the naming convention established above for their respective pipelines.

5 Experimental Results

In this section, we discuss the logistical settings of our experiments and discuss the performance of our baseline models on PGNSC.

5.1 Experimental Setup

Data splits. We divide the data according to an 80:10:10 split for the training, validation, and testing sets, respectively. Each experimental number shown is the averaged result of 10 experiments.

Evaluation metrics. We use the GMAUC metric (geometric mean of AUC and AUPR) as the primary evaluation metric in our work to more aptly capture performance on both positive and negative edge predictions, as shown in prior work (Poursafaei et al.). We also follow the conventions set by prior link prediction tasks and include AUPR and AUC as our evaluation metric (You et al., 2020; Kumar et al., 2020; Wu et al., 2022), which are available in the Appendix.

Experiments. We perform a grid search across the performances of five community aggregation methods and seven node embedding methods to investigate the impact of our proposed node representation methods and its interaction with community formulation. The five community aggregation methods are Biased-Random assignment, Unbiased-Random assignment, Engagement (Taylor et al., 2023), Interaction (Komorowski et al., 2018), and Influence-Passivity (Romero

et al., 2010). The seven node feature representation pipelines are Longformer, PRIMERA-BERT, GPT4-BERT, Llama2-BERT, PRIMERA-CLIP, GPT4-CLIP, and Llama2-CLIP.

5.2 Link Prediction Model

For the dynamic link prediction module, we use HTGNN (Heterogeneous Temporal Graph Neural Network) (Fan et al., 2021) as it is a state-of-the-art model for dynamic link prediction in a heterogeneous setting. HTGNN consists of multiple heterogeneous temporal aggregation layers, each employing hierarchical aggregation mechanisms that are combined to yield the spatio-temporal embedding for each node: intra-relation aggregation, inter-relation aggregation, and across-time aggregation.

Intra-relation Aggregation For each node v with type $\phi(v)$ at timestamp t, the feature vector $\mathbf{x}_v^t$ is projected into a unified feature space:

$$\mathbf{h}_{v,r}^{t,l} = \bigoplus_{u \in \mathcal{N}_r^t(v)}^{intra} \left(\mathbf{h}_u^{t,l-1}; \mathbf{\Theta}_{\text{intra}} \right), \quad (1)$$

$$\mathbf{h}_{v,R}^{t,l} = \bigoplus_{r \in R(v)}^{inter} \left(\mathbf{h}_{v,r}^{t,l}; \mathbf{\Theta}_{\text{inter}} \right), \quad (2)$$

$$\mathbf{h}_{v,ST}^{t,l} = \bigoplus_{1 \le t' \le T}^{across} \left(\mathbf{h}_{v,R}^{t',l}; \mathbf{\Theta}_{\text{across}} \right). \quad (3)$$

where $\bigoplus$ denotes the aggregation function, and the terms are defined as follows:

$\mathcal{N}_r^t(v)$ represents the set of neighbors of node v at timestamp t for relation type r, $R(v)$ denotes the set of relation types connected to node v, T is the total number of timestamps considered in the model. The final embedding for each node is the sum of its embeddings across all timestamps:

$$\mathbf{h}_v = \sum_{t=1}^{T} \mathbf{h}_v^{t,L}.$$

Finally, HTGNN is trained with cross entropy loss:

$$\mathcal{L} = \sum_{v \in \mathcal{V}_L} J(y_v, \hat{y}_v) + \lambda \|\mathbf{\Theta}\|_2^2, \quad \hat{y}_v = \sigma(\text{MLP}(\mathbf{h}_v)),$$

5.3 Overall Information Pathways Prediction Results

In this section, we will discuss the observations made and conclusions drawn from the experimental results shown in Figure 5.

Community Aggregation Method Comparison results. We observe that feature initialization pipelines (FI) show improvements of 6.55 to 10.98 on pathways mapped to communities constructed using the Influence-Passivity (IP) heuristic. We believe that IP benefits from factoring in both the influence exerted on outgoing edges and receptivity of users to incoming edges as opposed to Interaction and Engagement scores, which simply compare the relative influence exertion of users.

We also find that the Engagement community aggregation methods yields similar results to Biased Random community assignment, which indicates that heuristics relying on user-level network structure (Interaction and IP) are able to more meaningfully characterize communities than can engagement metrics. These results support the assertion that relying on the frequency of interactions provides more insight into user behavior than aggregated user engagement metrics (Romero et al., 2010).

Feature Initialization Method Comparison results. We observe that FI pipelines encoded by BERT offer an average 1.1 point improvement over those using CLIP across all community aggregation methods; this further increases to 1.6 point average improvement when considering the non-random aggregation heuristics. The two-stage FI pipelines provide an average improvement of 5.4 points over one-stage FI pipelines across all community heuristics, which increases to 6.2 points when excluding randomized community heuristics. We believe the reason behind the discrepancy between the effectiveness of BERT encodings and multimodal CLIP encodings is two-fold. First, CLIP embeddings are smaller due to model output parameters. Second, a manual investigation yields that, while images enhance the representation of articles they accompany, they may not aid in differentiating one community from another (See Appendix for example).

We also observe a slight trend in Llama-2 summary modules outperforming those using Primera and GPT4-based summary modules, and Llama-2 has an average 1.2 point improvement over PRIMERA under non-random community aggregation heuristics. However, this difference shrinks to 0.8 points when random settings are included, and the average relative differences in performance between Llama-2 and GPT4 and between GPT4 and PRIMERA are similarly insignificant.

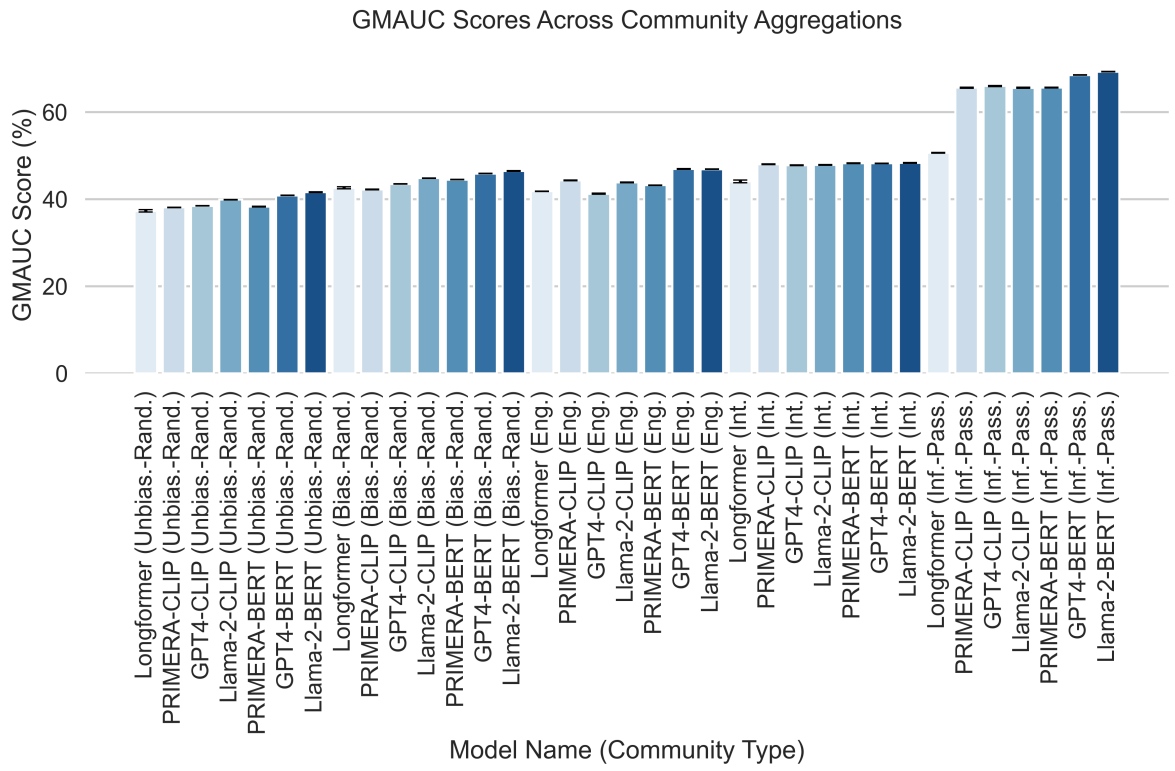

Figure 5: Overall Results. We provide the performance results of our baseline pipelines (GMAUC). Please see Appendix (Table 10) for full tabular representation including AUC and AUPR.

6 Related Works

6.1 Community-based Information Propagation Prediction on Social Networks

There are many existing methods designed to improve the performance of link prediction over social media network representations (Daud et al., 2020; Kumar et al., 2020; Wu et al., 2022; Haghani and Keyvanpour, 2019; Toprak et al., 2023). Several recent works have framed the link prediction problem as predictions of information propagation (Taylor et al., 2023; Jin et al., 2023). These works both define information pathways as the trajectory of individual pieces of information through communities of users.

Prior work made use of Reddit data, equating user participation in Subreddits with community membership and predicting information propagation (Jin et al., 2023). Many social networks that generate information cascades do not have explicit communities, thus Taylor et al. developed the Engagement heuristic to identify implicit communities using engagement metrics (Favorites, Retweets, etc.) on Twitter COVID-19 data (Taylor et al., 2023). Using this heuristic, Taylor et al. showed that using collections of articles to construct the feature representations of these communities yielded better static link prediction results over using tweet histories to represent individual users.

Our work continues in this vein and uses four distinct community construction methodologies (including the Engagement heuristic) and novel LLM-enhanced node feature generation methods for community feature representations (Taylor et al., 2023).

6.2 Multi-Modal Knowledge Graph Representations

Multi-modal graph representations have recently received increased attention because they more fully represent problems and lead to improved performance on downstream tasks (Peng et al., 2023; Liang et al., 2021; Zheng et al., 2022; Wei et al., 2019; Ektefaie et al., 2023). These improvements have been observed in diverse applications, from misinformation detection to transport systems (Abdali, 2022; Zhang et al., 2023; Tian et al., 2022; Rahmani et al., 2023; Jin et al., 2023; Chen et al., 2023a).

The recent revolution of LLMs has inspired re-

cent work to investigate the potential benefits of incorporating LLMs into graph neural networks (He et al., 2023; Chen et al., 2023b). Further works has shown that it is possible to use zero-shot LLM summarization in concert with graph embeddings to predict the news consumption habits of individual users (Chen et al., 2023a; Liu et al., 2023). These successes inspired prior work to use a longform text encoder to construct the node feature representations of communities based on collections of articles to predict information propagation (Taylor et al., 2023; Beltagy et al., 2020).

Our work furthers this direction by using SOTA LLMs to generate and encode community embeddings of a highly-curated, novel dataset. We demonstrate that downstream link-prediction performance benefits significantly from LLM-enhanced community summarization and outperforms prior approaches on our dataset.

7 Conclusion

In this work, we present a novel benchmark for dynamic link prediction on heterogeneous graphs in the novel application domain of information pathway prediction. We consider 5 community aggregation heuristics and 7 different feature initialization pipeline constructions, all of which show the improvement that LLM-enhanced feature representations can have on downstram link prediction. We also show that the community construction method significantly impacts overall performance, and our results support prior assertions of the benefits of considering interaction network structure (Romero et al., 2010).

It is our hope that PGNSC will make LLM-enhanced graph tasks more accessible, fill the necessity for a large dynamic heterogeneous graph dataset, and last, but not least, offer a challenging setting for incorporating community aggregation into information propagation prediction tasks.

Limitations

This work presents a novel benchmark, 5 community aggregation heuristics, and 7 feature initialization pipelines. While we mentioned scalability issues of existing methodologies, we acknowledge the simplicity of the community aggregation heuristics included in this work. We also acknowledge the limitations of our data: as mentioned in the main paper, many response tweets are missing engagement metrics, meaning that the Engagement and Interactions scores are being computed over incomplete data. Lastly, we have included the sources used to determine prominent global news sources, but acknowledge the bias inherent in the process of deciding whether to include news sources as communities or when excluding countries from data collection.

Ethics Statement

Our benchmark is intended to encourage exploration in the information propagation domain while avoiding any leakage of private information. The data we use in this work has been fully de-identified and only uses numerical codes to refer to users. Though it is very difficult to reconstruct protected personal information from such data, there remains some small risk that future models may be able to do so. The dataset we intend to release contains only de-identified user identification numbers to denote community membership and edges contain the engagement metrics of the associated tweet; no other metadata has not been retrieved. We have taken all known steps to prevent private information leakage.

Acknowledgements

This research was supported in part by the Defense Advanced Research Projects Agency (DARPA) under grant numbers HR00112290103 and HR0011260656.

We extend our heartfelt gratitude to Maximillian Sloan, Alvyn Wang, Xiaofu Ding, Nuan Wen, and Jiaying Li for their contributions to the implementation of our methods. Special thanks are due to Arjun Pawar, Xinran Feng, and Jacob Boughter for their diligent efforts in collecting and organizing background information on the various news sources referenced in this study. Additionally, we are immensely thankful to Dr. Dana Watson Cairns for her expert advice and constructive feedback on our manuscript.

The contributions of everyone involved have been pivotal in the advancement of this research, and we are deeply appreciative of their efforts.

References

Audit bureau of circulations india. Audit Bureau of Circulations. Used for determining inclusion of news sources.

a. Bbc guide for india. BBC. Used for determining inclusion of news sources.

b. Bbc media guide brazil. BBC. Used for determining inclusion of news sources.

China central television (cctv) reports. CCTV. Used for determining inclusion of news sources.

Digital news report. Reuters Institute for the Study of Journalism. Used for determining inclusion of news sources.

Eurotopics. https://www.eurotopics.net. Used for determining inclusion of news sources.

Foundations and Modeling of Dynamic Networks Using Dynamic Graph Neural Networks: A Survey | IEEE Journals & Magazine | IEEE Xplore.

Mass media in russia. Used for determining inclusion of news sources.

Media bias/fact check. mediabiasfactcheck.com. Used for assessing news sources.

News agencies - russia. Used for determining inclusion of news sources.

Oxford studies. University of Oxford. Used for determining inclusion of news sources.

Pew research center reports. Pew Research Center. Used for determining inclusion of news sources.

Sara Abdali. 2022. Multi-modal misinformation detection: Approaches, challenges and opportunities.

Khaled Ammar and M. Tamer Özsu. 2014. WGB: Towards a Universal Graph Benchmark. In *Advancing Big Data Benchmarks*, Lecture Notes in Computer Science, pages 58–72, Cham. Springer International Publishing.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Ferenc Béres, Róbert Pálovics, Anna Oláh, and András A. Benczúr. 2018. Temporal walk based centrality metric for graph streams. *Applied Network Science*, 3(1):1–26. Number: 1 Publisher: SpringerOpen.

Hao Chen, Runfeng Xie, Xia Cui, Zhou Yan, Xin Wang, Zhanwei Xuan, and Kai Zhang. 2023a. Lkpnr: Llm and kg for personalized news recommendation framework. *ArXiv*, abs/2308.12028.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. 2023b. Exploring the potential of large language models (llms) in learning on graphs.

Nur Nasuha Daud, Siti Hafizah Ab Hamid, Muntadher Saadoon, Firdaus Sahran, and Nor Badrul Anuar. 2020. Applications of link prediction in social networks: A review. *Journal of Network and Computer Applications*, 166:102716.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. 2023. Multimodal learning with graphs.

Yujie Fan, Mingxuan Ju, Chuxu Zhang, Liang Zhao, and Yanfang Ye. 2021. Heterogeneous temporal graph neural network. *ArXiv*, abs/2110.13889.

Naomi Forman-Katz. 2022. News platform fact sheet.

Alessio Gravina and Davide Bacciu. 2023. Deep learning for dynamic graphs: models and benchmarks. ArXiv:2307.06104 [cs].

Derek Greene, Dónal Doyle, and Pádraig Cunningham. 2010. Tracking the Evolution of Communities in Dynamic Social Networks. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 176–183.

Sogol Haghani and Mohammad Reza Keyvanpour. 2019. A systemic analysis of link prediction in social network. *Artificial Intelligence Review*, 52(3):1961–1995.

Xiaoxin He, Xavier Bresson, Thomas Laurent, and Bryan Hooi. 2023. Explanations as features: Llm-based features for text-attributed graphs.

Sameera Horawalavithana, Ellyn Ayton, Anastasiya Usenko, Shivam Sharma, Jasmine Eshun, Robin Cosbey, Maria Glenski, and Svitlana Volkova. 2022. EXPERT: Public Benchmarks for Dynamic Heterogeneous Academic Graphs. ArXiv:2204.07203 [cs].

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. 2023. Temporal Graph Benchmark for Machine Learning on Temporal Graphs. ArXiv:2307.01026 [cs].

Yiqiao Jin, Yeon-Chang Lee, Kartik Sharma, Meng Ye, Karan Sikka, Ajay Divakaran, and Srijan Kumar. 2023. Predicting information pathways across online communities. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Marlen Komorowski, Tien Do Huu, and Nikos Deligiannis. 2018. Twitter data analysis for studying communities of practice in the media industry. *Telematics and Informatics*, 35(1):195–212.

Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. 2020. Link prediction techniques, applications, and performance: A survey. *Physica A-statistical Mechanics and Its Applications*, 553:124289.

Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 933–943. ArXiv:1803.03697 [cs].

Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. 2008. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110. Publisher: American Physical Society.

Pengxiang Li, Hichang Cho, and Yuren Qin. 2022. (in)consistency matters: An account of understanding the perception of inconsistent expressions on social media. *Frontiers in Psychology*, 13.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4707–4715, New York, NY, USA. Association for Computing Machinery.

Jacob Liedke. 2022. U.s. adults under 30 now trust information from social media almost as much as from national news outlets.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2023. A first look at llm-powered generative news recommendation. *ArXiv*, abs/2305.06566.

Tore Opsahl. 2011. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. ArXiv:1006.0887 [physics].

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking : Bringing order to the web. In *The Web Conference*.

George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. 2021. Transfer Graph Neural Networks for Pandemic Forecasting. ArXiv:2009.08388 [cs, stat].

Jinghui Peng, Xinyu Hu, Wenbo Huang, and Jian Yang. 2023. What is a multi-modal knowledge graph: A survey. *Big Data Research*, 32:100380.

Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards Better Evaluation for Dynamic Link Prediction.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Saeed Rahmani, Asiye Baghbani, Nizar Bouguila, and Zachary Patterson. 2023. Graph neural networks for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 24(8):8846–8885.

Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. 2010. Influence and passivity in social media. *Proceedings of the 20th international conference companion on World wide web*.

Giulio Rossetti. 2017. $\text{RD}\small{\text{YN}}$: graph benchmark handling community dynamics. *Journal of Complex Networks*, 5(6):893–912.

Naw Safrin Sattar and Shaikh Arifuzzaman. 2022. Scalable distributed Louvain algorithm for community detection in large graphs. *The Journal of Supercomputing*, 78(7):10275–10309.

Alexander K. Taylor, Nuan Wen, Po-Nien Kung, Jiaao Chen, Violet Peng, and W. Wang. 2023. Where does your news come from? predicting information pathways in social media. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yijun Tian, Chuxu Zhang, Zhichun Guo, Yihong Ma, Ronald Metoyer, and Nitesh V. Chawla. 2022. Recipe2vec: Multi-modal recipe representation learning with graph neural networks.

Mustafa Toprak, Chiara Boldrini, Andrea Passarella, and Marco Conti. 2023. Harnessing the power of ego network layers for link prediction in online social networks. *IEEE Transactions on Computational Social Systems*, 10(1):48–60.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1437–1445, New York, NY, USA. Association for Computing Machinery.

Haixia Wu, Chunyao Song, Yao Ge, and Tingjian Ge. 2022. Link prediction on complex networks: An experimental survey. *Data Science and Engineering*, 7(3):253–278.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Jiaxuan You, Rex Ying, and Jure Leskovec. 2020. Design space for graph neural networks. *ArXiv*, abs/2011.08843.

Jianping Zeng and Hongfeng Yu. 2018. A Scalable Distributed Louvain Algorithm for Large-Scale Graph Community Detection. In *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 268–278, Belfast. IEEE.

Ji Zhang, Leonard Tan, and Xiaohui Tao. 2019a. On relational learning and discovery in social networks: a survey. *International Journal of Machine Learning and Cybernetics*, 10(8):2085–2102.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019b. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *ArXiv*, abs/1912.08777.

Taohong Zhang, Saian Chen, Aziguli Wulamu, Xuxu Guo, Qianqian Li, and Han Zheng. 2023. Transgnet: transformer and graph neural network based multi-modal data fusion network for molecular properties prediction. *Applied Intelligence*, 53(12):16077–16088.

Shuai Zheng, Zhenfeng Zhu, Zhizhe Liu, Zhenyu Guo, Yang Liu, Yuchen Yang, and Yao Zhao. 2022. Multi-modal graph learning for disease prediction. *IEEE Transactions on Medical Imaging*, 41(9):2207–2216.

8 Appendices

8.1 Full Selected List of Dynamic Graph Benchmarks

Please see Table 8 for the full list of graph benchmarks used to inform the decisions made in this work.

Training and evaluation. We select the best epoch based on the highest GMAUC score of the link prediction task on the validation set.

Frameworks. Our entire codebase is implemented in PyTorch.[1] The implementations of the transformer-based models are extended from the Huggingface[2] codebase (**?**).

8.2 Experiment Details

We report the median result for ten runs with different random seeds (by default). All the models in this work are trained on NVIDIA A100 GPUs on a Ubuntu 20.04.2 operating system.

8.3 LLM Community Summaries and Prompt Justification

We performed a manual tuning of 20 distinct prompt candidates and selected the prompt template that yielded the best results across the set of community summaries (150), which was a small enough set to manually validate.

Below is an example of the GPT4 response to the following prompt :"Summarize past content of NPR with respect to COVID-19 (Do not tell me you're an AI)"

"NPR has provided comprehensive coverage of COVID-19 since the early beginnings of the pandemic. Its content has ranged from daily updates on the number of COVID-19 cases, deaths and progress on vaccines to in-depth features and analyses. They've frequently reported on the impact of the virus on businesses and the economy, as well as on daily life, education, and mental health issues.

NPR has regularly reported on the response of federal, state, and local governments to the pandemic, including policies put in place and financial relief measures. This has also included coverage of controversy around stay-at-home orders, mask mandates, and businesses re-openings.

They've conducted numerous interviews with public health officials, researchers, doctors, and individuals affected by the pandemic. They've also widely covered the development of COVID-19 vaccines, including the process, distribution issues, and public perception.

Specifically, there have been stories on the effect of the pandemic on marginalized communities, including racial disparities in healthcare and its impact. NPR also looked at international aspects of the pandemic, from the initial outbreak in Wuhan, China to comparisons with other countries' responses and strategies.

NPR's COVID-19 coverage tends to focus on factual reporting and sharing human stories and includes input from experts in multiple fields. They've encouraged listeners to practice safety measures and have made an attempt to dispel myths around COVID-19. The tone of the coverage, though serious, often brings in elements of empathy, humanity, and sometimes humor to lighten the heavy subject matter."

8.4 Overall Model Results Table

Please see Table 10 for the full tabular representation of the experiments shown in Figure 5.

8.5 Image Relevance Discussion

To provide an example included in our manual error analysis, we consider the following article from Hindustan Times. The article concerns beach crowds during a time of heightened infections, but the images are only of 'beach crowds'. Further work in gauging image relevance is necessary to explore the full benefits of incorporating multimodal article information into the feature initialization pipeline.

[1] https://pytorch.org/

[2] https://github.com/huggingface/transformers

Table 8: Full collection of sampled graph benchmarks.

Name	#Nodes	#Edges	Seq. len.	Snapshot sizes (N/E)	Granularity	Type
Autonomous systems	7,716	13,895	733	10.6k-7.2k / 13.2k	daily	D
Bitcoin-OTC	3,783	24,186	24,186	–	seconds	C
Bitcoin-Alpha	3,783	24,186	24,186	–	seconds	C
CONTACT	5.3k	35.6k	38.4k	–	seconds	C
ENRON	151	2.2k	50.6k	–	unix timestamp	C
Elliptic	203.8k	234.4k	49	1.6k-2.2k / 1.2k-2.0k	49 steps	D
FB-Forum	899	7.1k	33.7k	–	unix timestamp	C
FB-Covid19	152 (ENG), 154 (FRA), 96 (ITA), 54 (ESP)	13.7k (ENG), 23.4k (FRA), 7.7k (ITA), 12.2k (ESP)	610 (ENG), 105 (FRA), 58 (ITA), 53 (ESP)	152/13.7k (ENG), 154/23.4k (FRA), 96/7.7k (ITA), 54/12.2k (ESP)	daily	D
HYPERTEXT09	113	2.5k	30.2k	–	seconds	C
IA-Email-EU	986	24.9k	332.3k	–	seconds	C
LastFM	2k	15.5k	1.3M	–	unix timestamp	C
Loop	207	2.4k	2.4k	207 / 2.4k	5 mins	S/T
METR-LA	207	1.5k	34.3k	207 / 1.5k	5 mins	S/T
Montevideo	675	690	740	675 / 690	hourly	S/T
MOOC	7.1k	411.7k	178.4k	–	unix timestamp	C
PDMS03	358	442	26.2k	358 / 442	5 mins	S/T
PDMS04	307	290	16.3k	307 / 290	5 mins	S/T
PDMS07	883	709	28.9k	883 / 709	5 mins	S/T
PDMS08	170	137	17.9k	170 / 137	5 mins	S/T
PDMSBay	325	2.4k	5.2k	325 / 2.4k	5 mins	S/T
PDMSBy	1.3k	19.1k	1.3M	–	5 mins	S/T
RADOSLAW	167	5.5k	82.9k	–	seconds	C
Reddit	11k	73.5k	672.4k	–	unix timestamp	C
Reddit Hyperlink Network	55.9k	858.5k	85.9k	–	unix timestamp	C
SBM-synthetic	1k	130.4k	50	1k / 933.1-105.4k	50 steps	S/T
SOC-Wiki-Elec	7.1k	103.7k	107.1k	–	unix timestamp	C
SZ-taxi	1.2k	1.5k	1.3k	156 / 532	15 mins	S/T
Traffic	4.4k	9k	2.2k	4.4k / 9k	hourly	S/T
Twitter-Tennis	1k	40.3k	1.9k	1k / 41-936	hourly	D
UCI messages	1.9k	20.3k	59.8k	–	unix timestamp	D
Wikipedia	9.2k	18.3k	157.5k	–	unix timestamp	C
tgbl-wiki	9.2k	157.5k	152.8k	–	–	D
tgbl-review	352.6k	4.9M	6.9k	–	–	D
tgbl-coin	638.5k	22.8M	1.3M	–	–	D
tgbl-comment	994.8k	44.3M	31M	–	–	D
tgbl-flight	18.1k	67.2M	1.4k	–	–	D
tgbn-trade	255	468.2k	32	–	–	D
tgbn-genre	1.5k	17.9M	133.8k	–	–	D
tgbn-reddit	11.8k	27.2M	21.9M	–	–	D
tgbn-token	61.8k	72.9M	2M	–	–	D

Table 10: Model Pipeline Performances

Pipeline Configuration			Performance Metrics					
Comm. Agg.	LLM	Encoder	AUC	Var	AUPR	Var	GMAUC	Var
Unbiased-Random	-	Longformer	57.66	1.79E-06	24.18	1.39E-04	37.34	1.22E-04
	PRIMERA		58.56	3.52E-06	29.58	1.43E-04	41.62	5.36E-05
	GPT4	CLIP	60.82	1.13E-05	27.49	1.05E-04	40.89	1.03E-04
	Llama-2		61.88	3.67E-06	23.72	5.15E-05	38.31	1.03E-04
	PRIMERA		60.94	6.84E-05	26.14	2.41E-04	39.91	4.95E-04
	GPT4	BERT	62.65	2.04E-05	23.66	1.43E-04	38.50	2.11E-04
	Llama-2		52.82	6.84E-05	27.54	2.05E-04	38.14	4.95E-04
Biased-Random	-	Longformer	57.86	2.58E-03	32.70	1.075E-02	42.63	4.68E-03
	PRIMERA		61.77	1.65E-05	28.89	4.07E-06	42.24	6.24E-06
	GPT4	CLIP	61.79	3.81E-06	30.64	8.65E-07	43.51	5.13E-08
	Llama-2		62.12	5.62E-07	32.38	7.43E-07	44.84	6.50E-07
	PRIMERA		63.34	2.70E-05	31.28	1.07E-05	44.51	5.60E-06
	GPT4	BERT	64.96	1.59E-05	32.43	2.19E-05	45.90	1.31E-05
	Llama-2		64.26	5.83E-06	33.66	5.84E-06	46.50	6.26E-06
Engage-ment	-	Longformer	65.17	1.16E-04	26.86	1.87E-07	41.84	1.09E-05
	PRIMERA		64.37	7.23E-05	30.56	4.20E-05	44.34	5.82E-06
	GPT4	CLIP	64.67	4.30E-06	26.38	3.77E-04	41.28	2.42E-04
	Llama-2		65.74	6.84E-05	29.26	1.86E-04	43.85	9.72E-05
	PRIMERA		64.20	1.13E-05	29.13	1.89E-06	43.24	2.02E-06
	GPT4	BERT	66.61	4.42E-05	33.06	8.62E-06	46.93	1.75E-05
	Llama-2		66.58	6.48E-06	33.02	9.08E-06	46.89	5.10E-06
Interaction	-	Longformer	63.43	2.38E-04	32.90	2.947E-02	44.11	1.247E-02
	PRIMERA		62.52	2.84E-05	36.94	9.92E-05	48.05	7.48E-05
	GPT4	CLIP	62.17	6.13E-05	36.75	4.34E-05	47.80	3.62E-05
	Llama-2		62.22	2.18E-05	36.83	1.19E-04	47.87	3.72E-05
	PRIMERA		63.52	2.02E-05	36.71	6.82E-05	48.29	5.46E-05
	GPT4	BERT	63.16	2.75E-05	36.83	6.19E-06	48.23	2.60E-06
	Llama-2		62.98	4.01E-05	37.06	7.62E-05	48.32	3.48E-05
Influence-Passivity	-	Longformer	66.85	3.99E-06	38.42	4.23E-04	50.66	1.69E-04
	PRIMERA		81.06	2.20E-05	53.12	3.66E-04	65.61	1.87E-04
	GPT4	CLIP	78.89	8.06E-05	55.25	3.07E-04	66.01	1.58E-04
	Llama-2		79.33	2.31E-04	54.25	3.93E-04	65.59	2.52E-04
	PRIMERA		78.95	7.15E-06	54.63	1.13E-04	65.67	3.22E-05
	GPT4	BERT	82.43	1.47E-05	56.92	2.43E-05	68.50	1.98E-06
	Llama-2		82.54	4.26E-06	58.17	2.02E-06	69.30	2.73E-07

8.6 List of countries and News Sources

We include a list of the news sources included in this work in Table 11, as well as the resources that we used to compile the list of news sources used in this work shown in Table 12.

Table 11: News Sources by Country with Selected URLs

Country	News Source	Selected URLs
US	Fox, CNN, NPR, NBC, ABC, CBS News, MSNBC, nyt, Facebook, The Washington Post	https://www.foxnews.com, https://www.cnn.com, https://www.nytimes.com, https://www.washingtonpost.com
Russia	Russia Today, Sputnik, TASS, Interfax, Ria Novosti, Argumenty i Fakty, The Moscow Times	https://www.rt.com, https://sputniknews.com, https://tass.ru
China	China Media Group, CGTN, People's Daily, Xinhua News Agency, China News Service, China Daily	https://www.cgtn.com, https://news.cn, https://www.chinadaily.com.cn
Hong Kong	TVB News online, Hk01, Now TV News online, Headline Daily online, Oriental Daily News online	https://news.tvb.com, https://www.hk01.com
Taiwan	ETtoday online, TVBS News online, Line News, EBS News online, Sanlih E-TV News	https://www.ettoday.net, https://news.tvbs.com.tw
Britain	BBC News online, Guardian online, Sky News online, ITV news, Channel 4 News, Daily Mail/Mail on Sunday	https://www.bbc.com, https://www.theguardian.com
France	France Televisions, BFM TV News, TF1 News, M6 News, 20 minutes online, France Info	https://www.francetelevisions.fr, https://www.bfmtv.com

Table 11 – continued from previous page

Country	News Source	Selected URLs
India	Press Trust of India, CNN News 18, Times of India, The Hindu, Indian Express, Republic TV/Republic World, Hindustan Times, NDTV, Dainik Jagran, Dainik Bhaskar, Hindustan, Amar Ujala, Malayala Manorama, Sakal, Dina Thaanthi, Eenadu, Lokmat	https://www.ptinews.com, https://timesofindia.indiatimes.com
Brazil	Globo News, UOL online, Record News, O Globo, Band News, CNN Brazil, O dia, Folha de Sao Paulo, O Estado de Sao Paulo, Rio Times	https://g1.globo.com, https://www.uol.com.br
Turkey	Sozcu, CNN turk, Haberturk, TRT news, Sondakika, Mynet, NTV, Hurriyet, Cumhuriyet	https://www.sozcu.com.tr, https://www.cnnturk.com
Italy	La Repubblica, Corriere della Sera, Il Sole 24 Ore, TGCom 24, ANSA, Notizie Libero, Il Fatto Quotidiano	https://www.repubblica.it, https://www.corriere.it
Spain	El Pais, OKDiario, El Mundo, La vanguardia, ABC Spain, La Razon, Antena 3, Telecinco, RTVE	https://elpais.com, https://www.elmundo.es
Argentina	Telefe News, Todo Noticias, Canal 13 News, C5N, A24, Infobae, Clarin online, La Nacion online	https://noticias.mitelefe.com, https://www.infobae.com

Table 11 – continued from previous page

Country	News Source	Selected URLs
Colombia	El Tiempo, Noticias Caracol, Las 2 Orillas, El Espectador, Pulzo, Noticias RCN, Noticias uno, Q hubo, El nuevo siglo, City Paper Bogota, Colombia Reports	http://www.eltiempo.com, https://noticias.caracoltv.com
Mexico	TV Azteca News, Televisa News, El Universal, Imagen News, Milenio Noticias, UnoTV news online, Aristegui News	https://www.tvazteca.com, https://www.televisa.com
Israel	Israel Hayom, Maariv, Haaretz, Yedioth Ahronoth	https://www.israelhayom.com, https://haaretz.com
Iran	Tehran Times, Iran Daily, Iranian Students News Agency, Financial tribune, Resalat, Hamshahri, Kayhan, Ettelaat	https://www.tehrantimes.com, https://en.isna.ir
Germany	ARD News, ZDF News, RTL News, t-online, Spiegel online, Bild.de	https://www.tagesschau.de, https://www.zdf.de
Poland	TVN News, RMF FM / RMF24, Polsat News, Radio Zet, Fakt, Gazeta Wyborcza, Onet, WP, Interia	https://tvn24.pl, https://www.rmf24.pl

Table 12: Sources for Determining Inclusion of News Sources

Source Name	Link
Pew Research Center Reports (pew)	N/A
News Agencies - Russia (new)	N/A
Mass media in Russia (Wikipedia) (mas)	https://wikipedia.org
China Central Television (CCTV) Reports (cct)	N/A
Oxford Studies (oxf)	N/A
Digital News Report (Reuters Institute) (reu)	https://reutersinstitute.politics.ox.ac.uk/digital-news-report
BBC Guide for India (bbc, a)	N/A
Audit Bureau of Circulations India (aud)	N/A
Potential Political Leanings (Twitter, Reddit) (?)	Twitter, Reddit
Media Bias/Fact Check (med)	https://mediabiasfactcheck.com
BBC Media Guide Brazil (bbc, b)	N/A
Eurotopics (eur)	https://www.eurotopics.net