Where Does Your News Come From? Predicting Information Pathways in Social Media

ABSTRACT

As social networks become further entrenched in modern society, it becomes increasingly important to understand and predict how information (e.g., news coverage of a given event) is propagated across social media (i.e. information pathway), which helps the understandings of the impact of real-world information. Thus, in this paper, we propose a novel task, Information Pathway Prediction (IPP), which depicts the propagation paths of a given passage as a community tree (rooted at the information source) on constructed community interaction graphs where we first aggregate individual users into communities formed around news sources and influential users, and then elucidate the patterns of information dissemination across media based on such community nodes. We argue that this is an important and useful task because, on one hand, community-level interactions offer more stability than those at the user level; on the other hand, individual users are often influenced by their community, and modeling community-level information propagation will help the traditional link-prediction problem. To tackle the IPP task, we introduce LIGHTNING , a novel content-aware link prediction GNN model and demonstrate using a large Twitter dataset consisting of all COVID related tweets that LIGHTNING outperforms state-of-the-art link prediction baselines by a significant margin.

CCS CONCEPTS

• **Networks** → *Network performance analysis*; **Network performance modeling**; **Network experimentation**.

KEYWORDS

Machine Learning, Graph Neural Networks, Data Mining, Social Networks

ACM Reference Format:

. 2023. Where Does Your News Come From? Predicting Information Pathways in Social Media. In *Proceedings of The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Because of the revolution of social networks and online technology, pieces of information can spread uncontrollably, and in some cases, significantly affect modern society. This makes measuring how new information propagates from off-line real-world events through social media networks to trigger varied responses across communities of individuals critical to developing response strategies. We aim to predict the *information pathway* from its source to potential targets based on the type of sources (i.e., news organization, influencer, political alignment, social credibility, etc.), which can provide insights on understanding potential social impacts of our real world events. Besides measuring social impacts, elucidating information pathways at this scope has many downstream applications. For instance, understanding the pathway patterns that misinformation tends to follow would hint the scrutiny of the information following similar patterns to prevent the spread of misinformation.

Prior works studying information pathways mainly focus on utilizing information transition between *individual users* inside one single community for *structural analysis* [5, 9, 10, 18]. Despite some success, using only user-level structure often impacts the scalability of the models. These approaches also focus on user interactions within a single community of users and thus may have limited ability to generalize to new domains/communities. These challenges are compounded when considering multiple communities of users that have many cross-community interactions. We aim to overcome these limitations by *predicting* the information pathway between *groups of users (communities)* at the community-level and propose a novel prediction task to model the information propagation: Information Pathway Prediction (IPP), which consists of a series of link predictions, each representing one step of information propagation across communities and, when combined together, forming a tree structure rooted in the information source (see Figure 1.).

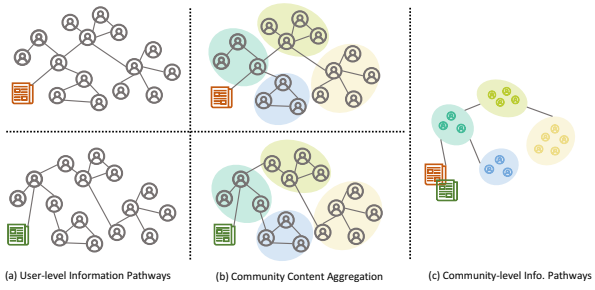

Figure 1: Example showing that coarsening the top user-level information pathway (IP) results in the same overall community-level IP as the bottom IP since the alternate user interaction occurred between the same two communities as the original user interaction.

There are several key challenges when designing and studying the IPP task over graphs representing the interactions of millions of users from different communities. First, extracting patterns of information dispersal is difficult when conducting link prediction over graphs consisting of millions of users with irregular posting schedules [11]. Second, only leveraging node and graph structure features does not fully incorporate the available information provided by the individual messages in information pathway instances.

SIGIR'23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

To address the first concern, we propose to coarsen the meta-graph of all information pathways by aggregating individual users into explicit *communities* of users to incorporate the influence of the implicit communities on individual users into our representations. We follow the definition of communities as self-organizing, voluntary, and open participation systems that are created and sustained through computer-mediated communications and centered around influencers and offline news sources [7, 8], which are referred to as community centers. Users that interact with source information from community centers are ranked by their measured influence with respect to that community center, and we define community members as users whose influence score for a given community center is greater than that same users influence scores for all other community centers. We displays an example of the information pathway in Figure 1, where each user has been mapped to their respective community center. To address the second concern, we propose Lightning , a content-aware link prediction model to model and identify the most likely information propagation pathways that incorporate message content into the edge representations.

Our main contributions are summarized as follows:

- We propose the novel Information Pathway Prediction Task which studies the information propagation on the community-level.
- We propose a novel content-aware link prediction model, Lightning , that demonstrably outperforms state-of-the-art link prediction baselines on the IPP task.
- We introduce a graph dataset constructed from COVID-19 Twitter data to study the IPP task and demonstrate the performance of our proposed Lightning model.

2 RELATED WORK

2.1 Link Prediction over Social Networks

There are many existing methods designed to improve the performance of link prediction over social media network representations [4, 12, 15]. However, these works are restricted to improving the understanding of the underlying structure of the social network[4, 17], including approaches that construct networks from communications between communities of nodes [13]. Our approach incorporates both the structure of the network connections and the text-based content they communicate. The additional information we include allows for a more expressive link prediction model that is able to capture the additional context-based dependencies unavailable to models solely relying on structural information.

2.2 Information Diffusion in Social Networks

Recent works have investigated how information propagates through social networks [5, 9, 10, 18]. These works illustrate how individual user messages influence the structure and dynamics of an information pathway only on the user-level[5], and the impact of the messaging behavior of influential users and groups of users with common topical interests [10, 18]. In contrast, our method of community aggregation incorporates both communities of users and influential individual users with text-based representations of past user messages and available offline content. Further, we establish Lightning , a GNN-based, content-aware method for modeling and predicting the diffusion of information.

3 INFORMATION PATHWAY PREDICTION

In this section, we'll described the problem definition of Information Pathway Prediction(IPP), the construction of information pathway representation and our proposed Lightning model.

3.1 Task Definition

Node Types	
Information Source	Article written by a News Org. or a message posted by an Influencer
Community	Set of users aggregated around community centers
Edge Types	
Propagated_from	*Community −> Community*: Represents a user from one community interacting with a message authored by a user in a different community.
Mentioned_by	*Information Source −> Community*: Indicates that a community authored a message containing an Information Source, starting an information pathway.
Written_by	*Information Source −> Community*: Indicates the community that is original author of the Information Source.

Table 1: Definition of IIP graph instances.

We want to perform Information Pathway Prediction (IPP), which consists of a series of link predictions that can be combined to form a tree(graph) structure rooted in the information source Figure 1. IPP is performed over information pathway instances (IPI), which are constructed by assigning each user from the social media graph to a given community (see subsection 4.2 for further details on aggregation). Communities consist of groups of users tied to a source of information (Community Center); in this task, we define Community Centers as either offline News Organizations (e.g., BBC, Fox News, etc.) or influential individual users (Influencers). IPIs consist of the community-level interactions in social media platforms that are triggered by a single message containing an Information Source. We define *Information Source* as either an article written by a News Organization or a message posted by an Influencer (see Table 1). We define both *Information Source* and *Community* as node types in IPIs. We define three common interactions among Communities and Information Sources as edge types (see Table 1). A community can interact with another community by sharing information(Propagated_from). For each *Information Source*, it can either be written by (Written_by) or mentioned by (Mentioned_by) a *Community*. For example, a news article from The Guardian is *Written_by* The Guardian, and if Reuters refer to this news article in its Twitter post, it's *Mentioned_by* Reuters.

3.2 Contextualizing Information Pathways

We aim to capture the dynamics of the context of information pathways content produced by each Community Center in different time steps. Providing context to the Community embeddings allows the downstream model to incorporate dynamic knowledge of topics that each community associates with, which can be used to identify similar Communities. To accomplish this, we leverage long sequence document-level modeling [1, 3] via the pre-trained

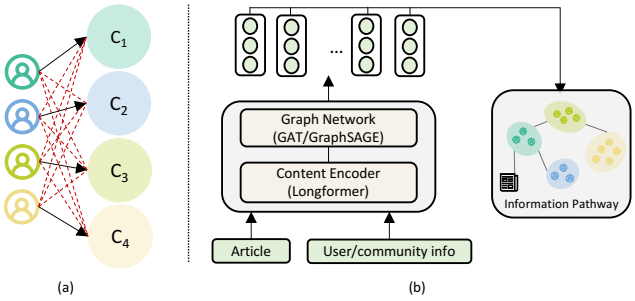

Figure 2: Figure 2a. illustrates users being assigned to the communities with a black edge indicating the highest engagement score. Figure 2b. illustrates the architecture of LIGHTNING.

Longformer model [2], which has shown consistent performance on various benchmarks [2], to derive content embeddings for each Community Center. To accurately represent the content associated with each Community Center, we aggregated all Information Sources authored by each Community Center and the individual community members from the time windows chosen for our experiments. The concatenation of all of these documents is used as input for Longformer for each Community Center, which outputs the embedding that we use as the set of node features for the Community node.

3.3 LIGHTNING

We formulate LIGHTNING as a link prediction model that incorporates the context provided by the content-aware feature for all Communities involved in information pathway instances into its predictions.

3.3.1 Node initialization. As mentioned in the previous section, we initialize our Longformer encoder $F_c(\cdot)$ and provide it the aggregation of n Information Sources to produce l-dimensional content-aware node embeddings.

$$\{h^c_{i,0}, ..., h^c_{i,l}\} = F_c(\{x_{i,0}, ..., x_{i,n}\})$$

We use $\mathbf{h}_i$ as the feature initialization for the i-th community node.

3.3.2 GNN Model. Following [16], we define a GNN model consisting of several message passing layers, where the k-th GNN layer is defined as:

$$\mathbf{h}_i^{(k+1)} = \text{SUM}\left(\text{RELU}\left\{\text{DROPOUT}(\mathbf{W}^k\mathbf{h}_j^k + \mathbf{b}^k), j \in \mathcal{N}(i)\right\}\right)$$

where h_j is the k-th layer embedding of node j, $\mathbf{W}^k$, $\mathbf{b}^k$ are trainable weights, and $\mathcal{N}(i)$ is the local neighborhood of node i.

3.3.3 Loss. During training, we seek to optimize the binary cross-entropy loss function.

$$\ell(x, y) = \text{MEAN}(L) = \{l_1, \ldots, l_N\}^\top$$

$$l_n = -w_n\left[y_n \cdot \log\sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))\right]$$

4 EXPERIMENTS

4.1 Data Collection

4.1.1 Twitter Data. We utilized the Twitter API to retrieve data from May 2020 to April 2021. From the retrieved data, we selected the 15-day period of time with the highest amount of activity (measured by number of tweets), which is 05/15/2020-05/30/2020. The 15-day period of data includes roughly 640 million tweets, 5.3 million distinct users, and 75,000 distinct English articles by selected news sources. We identify "source tweets" from this set that contain the URLs and titles of articles written by news sources or have been authored by influential users and retrieve both the text and following engagement metrics (defined in Section 4.2) for all reply tweets and retweets written in response to source tweets.

4.1.2 News Data. We use the URLs and article titles retrieved from the 15-day period of Twitter data (roughly 150K) to retrieve the content of the articles they refer to. We first filtered non-English news by removing article titles with non-ASCII characters, which left roughly 90K news titles. Next, we used the open-source package **GNews** to crawl the news content from Google News. This resulted in the full retrieval of roughly 75K news articles, each with their respective *Source Name*, *Source URL*, and *Article Content*. More statistics of the Twitter and News dataset are listed in Table 2.

Twitter data	
Number of Tweets	640 M
Number of Users	5.3 M
Number of News Link	144 K
Time Period	15 days
News data	
Number of News w/ Article	74.4 K
Number of Sources	5 K
Retrieve Rate	51.6%
Time Period	15 days

Table 2: Statistics of the Twitter and News data. For the News data, we retrieve news with articles from the news link presented in the Twitter data.

4.2 Data Pre-processing

Given the set of aggregated source tweets and their responses (tweet trees) we built above, we compute the Engagement score for each user in each tweet tree with respect to the community center (news source/influencer in the source tweet). When relying on the number of interactions with a 'community center', we find that the top users are typically bots designed to promote COVID news or individual user accounts with relatively small followings. Given that Twitter users broadcast their original tweets and retweets to their followers, we want to aggregate community members that receive the most engagement when propagating the news originating with a given community center. Thus, we constructed a user ranking score following the format outlined in [8] that uses the engagement metrics collected during tweet tree construction: number of favorites/retweets/replies, user IDs mentioned in the tweet, and Following/Follower ratio at the time of the tweet. The Engagement score illustrated in Equation 1 consists of the scaled

sum of tweet author statistics for user q: the first term represents the number of retweets user q has accrued for a given time window scaled by the sum of the number of retweets of all users; the other terms represent the equivalent formulations for replies, favorites, mentions, and following/follower ratio for the user at the given time window (Note: the follower following ratio is only retrieved for the latest day in the selected time window).

$$Engagement(q) = \frac{f_{\text{rtwt}}(q)}{\sum_{i=1}^{\|\mathcal{P}_{\text{rtwts}}\|} f_{\text{rtwts}}(x_i)} + \frac{f_{\text{replies}}(q)}{\sum_{i=1}^{\|\mathcal{P}_{\text{replies}}\|} f_{\text{replies}}(x_i)} + \frac{f_{\text{fvt}}(q)}{\sum_{i=1}^{\|\mathcal{P}_{\text{fvt}}\|} f_{\text{fvt}}(x_i)} + \frac{f_{@}(q)}{\sum_{i=1}^{\|\mathcal{P}_{@}\|} f_{@}(x_i)} + \frac{f_{f}ratio(q)}{\sum_{i=1}^{\|\mathcal{P}_{f}ratio\|} f_{f}ratio(x_i)} \quad (1)$$

We compute the Engagement Score for each user in each tweet tree with respect to the community center and assign users to the community for which they have the highest score (with ties being broken by random assignment).

4.3 Experimental Setup

4.3.1 Information Pathway Instances. To perform information pathway predictions, we use information pathway instances that range from 4 to 1000 pathway links and include up to 750 nodes. We also inject in one randomly sampled negative edge for each message present in the information pathway. To construct our community embeddings, we split the 15 days into 3 time windows: 05/16/20-05/20/20 was used to construct the community embeddings, 05/21/20-05/25/20 was used for training, and 05/26/20-05/30/20 was used for validation and testing. We use 2000 information pathways from the training time window and used 300 samples from the evaluation time for both validation and testing samples.

4.3.2 Model Settings. We create Information Source (IS) node embeddings for both articles written by News Organizations and original tweets posted by Influencers. To construct an IS node embedding for an article, we concatenate the article title, article content and the tweet text containing the article URL as input. For an Influencer tweet embedding, we only consider the text of the tweet as input. Given the input for the respective IS node, we use Longformer to generate a 768-dimensional embedding vector. For the Lightning setting, we construct content embeddings for 5,047 communities (aggregated from 1,612,254 users) using 34,722 articles and 54,866 influencer posts distributed across the Community Centers. To construct the Community embeddings, we concatenate all original tweets and include the number of replies, mentions and retweets of users within each community. We also aggregate each Community Center's Information Sources from the previous time window as input to Longformer to generate a 768-dimensional embedding vector. We also include a User Content embedding setting for the 25,844 user nodes present in the graph instances used for training and evaluation. The user node embeddings are generated in the same manner as Community node embeddings: we concatenate all original tweets and include the number of replies, mentions and retweets the user has gained in the previous time window as input to Longformer to to generate a 768-dimensional embedding vector. We implement Lightning using two standard GNN layers, **Graph Attention Networks** (**GAT**)[14] and **GraphSAGE**[6].

4.4 Baselines

According to the optimal settings outlined in [16] for the AmazonComputers link prediction dataset (identical graph structure to the IPP formulation), we benchmark the performance of GAT and GraphSAGE with randomly initialized embeddings (i.e. without content-aware embeddings at the Community or user-level). The baseline versions of GAT and GraphSAGE are implemented according to the optimal settings for the AmazonComputers link prediction dataset to provide a fair comparison. We also include a naïve "Always Negative" setting to provide additional context to our results. We follow the conventions set by prior link prediction tasks and use AUROC [12, 15, 16] as our evaluation metric.

4.5 Experimental Results

Table 3: Information Pathway Prediction results for GraphSAGE and GAT-based models. CL represents the performance of Community-level models and UL represents the performance of User-level models

Model	AUROC-CL	AUROC-UL
Always Neg.	50.00	50.00
GraphSage	83.27	73.43
GraphSage-Lightning	**86.83**	**77.01**
GAT	70.55	70.75
GAT-Lightning	**84.43**	80.53

Table 3 shows the overall results of our experiments. Overall, we find that the top performing models are GraphSAGE-Lightning and GAT-Lightning , supporting our assertion that content-aware community aggregations improves performance on the IPP task. When comparing the Lightning -versions to their respective community-level baseline, GAT-Lightning achieves an improvement of 14 points in AUROC score over the GAT community-level baseline. We also see a significant improvement (3 points AUROC score) by GraphSAGE-Lightning over the GraphSAGE community-level baseline. This again supports our claim that the content-aware embeddings stabilizes and improves performance of IPP. Lastly, when examining the effectiveness of community aggregation, we see that both GraphSAGE- and GAT-Lightning improve over their user-level baseline counterparts (with a 9 points and 4 points AUROC improvement). Overall, we find that content-aware community-level embeddings significantly improve the performance of link prediction models on the IPP task.

5 CONCLUSION AND FUTURE WORKS

In this work, we introduce the novel Information Pathway Prediction task. We provide an implementation of the task using COVID-19 Twitter Data and offline news sources, and provide content-aware link prediction model that shows the effectiveness of evaluating information pathways at the Community level as well as the benefits of encoding aggregated content for each community embedding. In the future, we plan to extend our Information Pathway Instances to incorporate other online social media networks too and improving upon our method of generating content-based embeddings.

REFERENCES

[1] Iz Beltagy, Arman Cohan, Hannaneh Hajishirzi, Sewon Min, and Matthew E Peters. 2021. Beyond paragraphs: NLP for long sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*. 20–24.

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

[3] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2978–2988. https://doi.org/10.18653/v1/P19-1285

[4] 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 (2020), 102716.

[5] Adrien Guille. 2013. Information diffusion in online social networks. In *SIGMOD'13 PhD Symposium*.

[6] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *NIPS*.

[7] Mandy Johnson, Chris Bledsoe, Jodi Pilgrim, and Hollis Lowery-Moore. 2019. Twitter: A Tool for Communities of Practice. *SRATE Journal* 28 (2019).

[8] 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 (2018), 195–212.

[9] Xiangjie Kong, Yajie Shi, Shuo Yu, Jiaying Liu, and Feng Xia. 2019. Academic social networks: Modeling, analysis, mining and applications. *Journal of Network and Computer Applications* 132 (2019), 86–103.

[10] Gueorgi Kossinets, Jon M. Kleinberg, and Duncan J. Watts. 2008. The structure of information pathways in a social communication network. In *Knowledge Discovery and Data Mining*.

[11] 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 (2020), 124289.

[12] 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 (2020), 124289.

[13] Anisha Kumari, Ranjan Kumar Behera, Bibudatta Sahoo, and Satya Prakash Sahoo. 2022. Prediction of link evolution using community detection in social network. *Computing* 104, 5 (2022), 1077–1098.

[14] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. 2017. Graph Attention Networks. *ArXiv* abs/1710.10903 (2017).

[15] Haixia Wu, Chunyao Song, Yao Ge, and Tingjian Ge. 2022. Link Prediction on Complex Networks: An Experimental Survey. *Data Science and Engineering* 7, 3 (2022), 253–278.

[16] Jiaxuan You, Rex Ying, and Jure Leskovec. 2020. Design Space for Graph Neural Networks. *ArXiv* abs/2011.08843 (2020).

[17] Ahmad Zareie and Rizos Sakellariou. 2020. Similarity-based link prediction in social networks using latent relationships between the users. *Scientific Reports* 10, 1 (2020), 20137.

[18] Hengmin Zhu, Xicheng Yin, Jing Ma, and Wei Hu. 2016. Identifying the main paths of information diffusion in online social networks. *Physica A: Statistical Mechanics and its Applications* 452 (2016), 320–328.