



Rumor Mitigation in Social Media Platforms with Deep Reinforcement Learning

Hongyuan Su
Department of Electronic
Engineering, BNRist, Tsinghua
University
Beijing, China

Yu Zheng
Department of Electronic
Engineering, BNRist, Tsinghua
University
Beijing, China

Jingtao Ding
Department of Electronic
Engineering, BNRist, Tsinghua
University
Beijing, China

Depeng Jin
Department of Electronic
Engineering, BNRist, Tsinghua
University
Beijing, China

Yong Li*
Department of Electronic
Engineering, BNRist, Tsinghua
University
Beijing, China

ABSTRACT

Social media platforms have become one of the main channels where people disseminate and acquire information, of which the reliability is severely threatened by rumors widespread in the network. Existing approaches such as suspending users or broadcasting real information to combat rumors are either with high cost or disturbing users. In this paper, we introduce a novel rumor mitigation paradigm, where only a minimal set of links in the social network are intervened to decelerate the propagation of rumors, countering misinformation with low business cost and user awareness. A knowledge-informed agent embodying rumor propagation mechanisms is developed, which intervenes the social network with a graph neural network for capturing information flow in the social media platforms and a policy network for selecting links. Experiments on real social media platforms demonstrate that the proposed approach can effectively alleviate the influence of rumors, substantially reducing the affected populations by over 25%. Codes for this paper are released at <https://github.com/tsinghua-fib-lab/DRL-Rumor-Mitigation>.

CCS CONCEPTS

• Computing methodologies → Reinforcement learning; Planning and scheduling.

KEYWORDS

rumor mitigation, social platforms, reinforcement learning

ACM Reference Format:

Hongyuan Su, Yu Zheng, Jingtao Ding, Depeng Jin, and Yong Li. 2024. Rumor Mitigation in Social Media Platforms with Deep Reinforcement Learning. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)*, May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3589335.3651556>

*Corresponding author (liyong07@tsinghua.edu.cn).



This work is licensed under a Creative Commons Attribution International 4.0 License.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0172-6/24/05.
<https://doi.org/10.1145/3589335.3651556>

1 INTRODUCTION

The rapid spread of rumors emerges as an unprecedented challenge, causing detrimental impact to public safety, health, and democracy. Central to the rapid dissemination of misinformation is the phenomenon of information cascades, where social network users influenced by the perceived information, adopt or share that information, leading to a cascading effect triggered and manipulated by strategically placed misinformation [2, 3]. Therefore, it is critical to develop an intelligent algorithm that can automatically synthesize rumor mitigation actions to counter the large number of rumors emerging everyday across the social network.

Current approaches for mitigating rumors follow two major paradigms. First, a subset of users are selected and suspended, *i.e.* removing nodes from the network to block the propagation of rumors [9]. Second, real information is introduced to the network, where several seed nodes are chosen for broadcasting real information to maximize its influence over the network [1]. However, these methods tend to be costly and cause significant annoyance to users. Therefore, in this paper, we focus on a new paradigm where rumors are mitigated with low business cost, and users are almost unaware of the mitigation actions [14]. We reserve all nodes and only block a small fraction of *links* in the social network to decelerate the information cascade of rumors, which induces minimal user disturbance and keeps information integrity.

Rumor mitigation via link deletion faces three main challenges. *First*, the candidate intervention options are extremely large as there are much more links than nodes in a social network, resulting in an enormous solution space. *Second*, it is difficult to capture the underlying mechanisms that drive the propagation of rumors due to its complexity and randomness. Thus, how to effectively leverage the domain knowledge of social networks remains largely unexplored by existing literature. *Third*, real-world social networks tend to exhibit diverse topologies and dynamics [12]. Consequently, it is necessary to develop an algorithm with sufficient generalization ability, to make it applicable in real rumor mitigation tasks.

To address these challenges, we develop a deep reinforcement learning (DRL) framework for rumor mitigation. To utilize the rich domain knowledge, we first extract critical features according to social network theories that capture the contribution of each node and edge in rumor propagation. We then design a graph neural

network (GNN) with link-route-aware message passing to capture local and global roles of links, respectively. A rumor-oriented policy network is further developed to select links for deletion, taking both rumor source and community structures into consideration. Our knowledge-informed agent embodies rumor propagation mechanisms, effectively promoting the accuracy and explainability of the generated rumor mitigation solutions. Finally, we design a randomized training environment that simulates social networks with diverse topologies and rumor propagation patterns, which guarantees the generalization ability of our approach.

To summarize, the contributions of this paper are as follows,

- We tackle the problem of rumor mitigation in a new paradigm with low cost and user disturbance, which only intervenes a minimal fraction of links of the social network.
- We develop a knowledge-informed agent and a randomized training algorithm to achieve generalizable rumor mitigation.
- Experiments on real-world social networks verify that our method reduces the influence of rumors substantially by over 25%.

2 PROBLEM STATEMENT

Consider a directed social network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ and $\mathcal{E}$ represent users and influence relationships (i.e. followers and followees) between them. Assume a rumor initiates from a source $s \in \mathcal{N}$, spreading through a diffusion model μ that leads to k nodes in the social network receiving the rumor. The impact of rumor can be defined as the percentage of the population affected,

$$\eta^\mu(\mathcal{G}, s) = \frac{\mathbb{E}_k[k | \mathcal{G}, s, \mu]}{|\mathcal{N}|}, \quad (1)$$

where $\mathbb{E}[x]$ denotes the expectation of the random variable x , and $|\mathcal{N}|$ represents the number of elements in the set $\mathcal{N}$. Rumor mitigation can be defined as selecting a subset of edges in the graph for removal to minimize the spread of rumors. The optimal edge set for deletion can be formulated as:

$$\mathcal{E}_D^* = \arg \min_{\mathcal{E}_D \subseteq \mathcal{E}, |\mathcal{E}_D|=d} \eta^\mu(\mathcal{G} \setminus \mathcal{E}_D, s), \quad (2)$$

where $\mathcal{E}_D$ represents the deleted edge set.

3 METHOD

Figure 1 illustrates the overview of the proposed framework where an agent sequentially removes edges on social networks to prevent rumors from spreading. A knowledge-informed model is designed to achieve decent edge selection, which encodes rich propagation-related features with a link-route-aware GNN model taking both rumor source and propagation dynamics into consideration. A rumor-oriented policy network is designed for edge-ranking, perceiving the hierarchical structure and the association with rumor source, enabling high-quality edge blocking for rumor mitigation. Moreover, we develop a randomized RL training approach to enhance the generalization ability of our model, addressing the challenge of diverse topological structure and uncertain rumor sources.

3.1 Knowledge-informed Edge Selection

Features for Topology and Propagation. Rumor spreading involves multiple aspects, such as rumor source, network structure, and the propagation mechanism. We design rich topological and

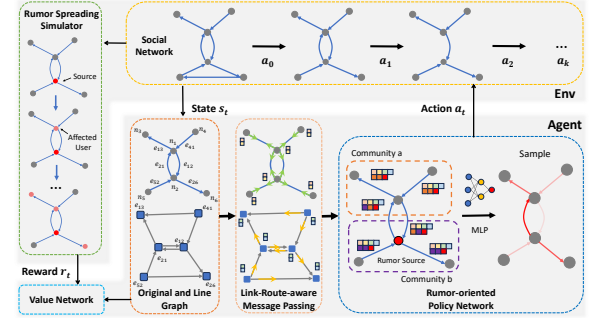


Figure 1: (Top) The environment provides observations of social network, controls the network based on the actions chosen by the agent, and returns the rewards through the rumor propagation simulations. (Bottom) Knowledge-driven edge selection which aggregates neighbors' information through link-route-aware message passing on both the original and the line graph using a GNN, and finally selects edges with the rumor-oriented policy network.

rumor spreading related features for nodes (FN) and edges (FE) in social networks, which contain the prior knowledge of rumor spreading and can be categorized into two types.

- *Topological features* include degree (FN1, FE6), centrality (FN2), and diffusion importance (FE7) [8], describing the topological structure of the social network and providing valuable insights into analyzing the spread of rumors. For example, diffusion importance indicates the efficiency of information propagation via edges, thus blocking edges with higher diffusion importance can efficiently prevent rumors from reaching more users.
- *Rumor propagation features* include shortest path length to rumor source (FN4), rumor source or not (FN5), and betweenness centrality (FN3, FE8) which are closely related to the characteristics of rumor propagation. It is evident that users close to the source tend to be affected by the rumor earlier. Therefore, edges closer to the source are expected to have a higher removal priority for rumor mitigation.

Notably, edge removal alters the topological structure of social network, thus these features are dynamically changing at each step.

Link-Route-aware Message Passing. Since edges represent influence relationships and serve as key components in rumor propagation, it is necessary to consider their multiple roles as both local user links and global propagation routes. Therefore, we propose a link-route-aware GNN to learn comprehensive edge embeddings, taking both aspects into consideration, as shown in Figure 1. We compute representations for nodes and edges within both the original and the dual line graph as follows,

$$\mathbf{n}_i^0 = \mathbf{W}_n^0 \mathbf{A}_{n_i}, \quad \mathbf{e}_{ij}^0 = \mathbf{W}_e^0 \mathbf{A}_{e_{ij}}, \quad (3)$$

$$\mathbf{n}_i^{l+1} = \mathbf{n}_i^l + \tanh \left(\sum_{e_{ij} \in \mathcal{E}, e_{ji} \in \mathcal{E}} \mathbf{W}_n^{l+1} \mathbf{n}_j^l \right), \quad (4)$$

$$\mathbf{e}_{ij}^{l+1} = \mathbf{n}_q^{*l+1} = \mathbf{n}_q^{*l} + \tanh \left(\sum_{e_{pq}^* \in \mathcal{E}^*} \mathbf{W}_e^{l+1} \mathbf{n}_p^{*l} \right), \quad (5)$$

$$\mathbf{e}_{ij}^f = [\mathbf{n}_i^l \parallel \mathbf{n}_j^l \parallel \mathbf{e}_{ij}^l], \quad (6)$$

Table 1: Basic properties of experimented online social networks. The Nodes and Edges columns indicate the maximum (average) number of corresponding elements in graph.

Network	Type	Nodes	Edges	Diameter
Twitter S	Directed	222 (78)	996 (528)	10
Twitter M	Directed	250 (171)	7990 (3086)	9
Twitter L	Directed	248 (215)	17930 (9847)	7
Facebook L	Undirected	1045 (417)	60050 (17018)	11

where $\mathbf{A}_{n_i}$ and $\mathbf{A}_{e_{ij}}$ are input attributes for nodes and edges, $\mathbf{W}_n^k$ and $\mathbf{W}_e^k$ are linear layers. Node and edge embeddings are obtained from primal and induced dual line graph, respectively, where edges on primal graph correspond to nodes on line graph, and the asterisk indicates elements in line graph. The output $\mathbf{n}_i^L$ and $\mathbf{e}_{ij}^L$ of the last propagation layer are the embeddings of nodes and edges, where L is a hyperparameter and $\parallel$ means concatenation.

By combining associated two node embeddings of each edge, $\mathbf{n}_i^L$ and $\mathbf{n}_j^L$, we can capture the influence relationships of local user links. Meanwhile, the edge embedding on the line graph, $\mathbf{e}_{ij}^L$, captures the global propagation route role of each edge. Thus the link-route-aware message passing effectively encode the roles of edges in connecting users and spreading rumors, providing valuable insights for the following edge selection.

Rumor-oriented Policy Network. We develop a rumor-oriented policy network that takes into account the edge embeddings from GNN, the rumor source, and the communities of social networks. As illustrated in Figure 1, the policy network calculates the scores of edges through a multi-layer perceptron (MLP) as follows,

$$C_i = \frac{1}{|C_i|} \sum_{n_k \in C_i} \mathbf{n}_k^L, \quad s_{ij} = \text{MLP}_P([\mathbf{e}_{ij}^L \parallel C_i \parallel C_j \parallel \mathbf{n}_s^L]), \quad (7)$$

where C_i donates the community that node i belongs to, and $\mathbf{n}_s$ donates the embedding of rumor source calculated by GNN. The agent then samples edges based on probabilities obtained by normalizing the edge scores from the policy network.

3.2 Generalized Model Training

As real-world social media platforms exhibit diverse characteristics in network topology and propagation dynamics, it is expected that a rumor mitigation algorithm to be generalizable across different scenarios. To achieve this, we introduce a randomized RL framework that allows the agent to learn effective rumor mitigation strategies from a large and diverse set of rumor propagation tasks. In specific, we train the agent with random network topology, where in each episode the environment generates a different social network. Meanwhile, for the same social network, we simulate the spread of rumors across various topological structures by randomly selecting different nodes as the rumor source at each training episode, which massively increase the number of training social networks.

4 EXPERIMENTS

4.1 Experiment Settings

Social networks Data. We conduct experiments based on the real-world online social networks including Twitter and Facebook from *Stanford Large Network Dataset Collection* [6]. Table 1 summarizes

Table 2: Rumor mitigation performance comparison.

Method	Twitter			Facebook L
	S	M	L	
HSD	16.12	13.48	12.93	13.34
HSC	18.96	20.95	18.65	25.00
GA	17.32	14.37	12.77	11.33
SA	19.30	13.93	13.34	10.98
PR	16.56	13.92	12.50	13.62
KED	19.80	22.16	17.57	18.30
GBP	28.53	26.21	24.32	28.46
DRLN	28.47	26.16	26.81	29.67
DRLE (ours)	34.81	31.22	32.18	37.26
impr% v.s. DRLN	+22.3%	+19.3%	+20.1%	+25.6%

the basic properties of the adopted datasets. For each network, we apply the susceptible-infectious-recovered (SIR) model [4] to simulate rumor propagation with $\gamma = 0.20$ and $\beta = 0.08$ [15].

Baselines and Evaluation. We include *traditional approaches* such as Pagerank (PR) [10], K-EDGEDELETION (KED) [13], and Greedy with Bond Percolation (GBP) [5]. We also compare our model with *heuristic search* approaches based on degree (HSD) and betweenness (HSB), and *evolutionary algorithms* including genetic algorithm (GA) [11] and simulated annealing (SA) [7]. Moreover, we include a recent DRL baseline [9] which selects nodes (DRLN) based on embeddings calculated from GNN.

4.2 Overall Performance

We remove 10% of the influencing relationships in the social network, while retaining at least 60% of the influencing relationships for each user. Table 2 illustrates the results of our model and baselines, and we have the following observations,

- **DRL-based methods have significant advantages over other baselines.** The DRL-based methods outperform HSC, KED and greedy method on diverse social networks. Specifically, they exceed Greedy method with an average improvement of 10.9%, 9.5%, and 19.5% in small, medium and large-sized social networks, respectively. Moreover, the DRL-based methods can be easily applied to complex social networks with stable outcome and reasonable computational cost.
- **Our proposed model achieves the best performance.** Our approach maximizes rumor mitigation across various social networks. Specifically, we observe a minimum improvement of 19.1% over suboptimal solutions. Notably, as the size of the social network increases, the performance of our approach in rumor mitigation further improves, achieving an over 37% reduction in rumor spread by blocking only 10% of the edges, and the improvement compared to other baselines rises from 19% to 25%.

We visualize the rumor propagation at each step in Figure 2. In the original network, rumor spreads rapidly upon their emergence, reaching the peak within only 10 steps and affecting more than 60% of the users in total. After implementing the proposed rumor mitigation strategy, there is a substantial suppression of rumors. Specifically, no more than 20% of users are affected in the first 10 steps of propagation, and the peak impact of rumors also dropped significantly from 35.6% to 10.1%. From a global perspective, our approach reduces the proportion of total users affected by rumors

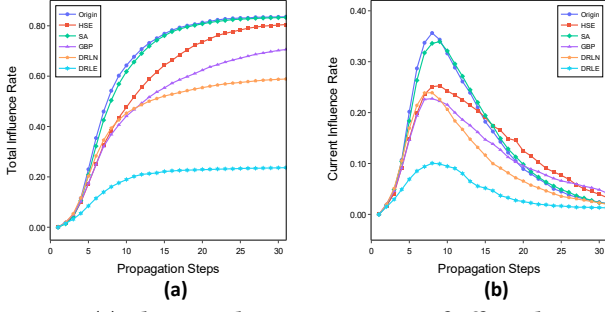


Figure 2: (a) The cumulative percentage of affected users at each propagation step. (b) The proportion of users who are spreading rumors at each propagation step.

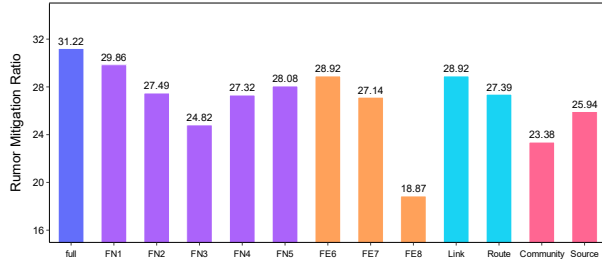


Figure 3: The result of ablation experiments, which removes features for nodes (FN) and edges (FE), GNN (Link and Route) and the rumor-oriented design in policy network (Community and Source), respectively. Best viewed in color.

from 83.6% to 23.7%, whereas baselines only temporarily block rumors, resulting in at least 58.8% of users ultimately being affected.

4.3 Ablation Study

We conduct ablation experiments to show the effectiveness of several key designs of the knowledge-informed model.

Features for Topology and Propagation Figure 3 shows the importance of each feature by removing them individually from the well-trained model. The removal of propagation-related edge centrality (FE8) results in the most significant performance degradation (-39.6%), and the decline is also significant (-20.5%) for propagation-related node centrality (FN3), which is reasonable as the propagation-related centralities identify critical edges for rumor propagation. The carefully designed features for topological structure and propagation can capture the key factor of rumor spreading, which is critical for rumor mitigation.

Link-Route-aware Message Passing The proposed GNN model combines the functions of edges as both links and routes in the spread of rumors. To assess the impact of propagation messages, we conduct experiments by eliminating message passing between nodes and edges, as illustrated in Figure 3. We can find that the removal of node and edge propagation results in a 7.4% and 12.3% reduction in the mitigation ratio, respectively. This result confirms the advantage of our link-route-aware message passing in aggregating local and global information.

Rumor-oriented Policy Network As shown in Figure 3, ablation experiments demonstrate the significance of community and rumor source identity in achieving high-quality edge selection. In specific,

when we neglect the correlation with the rumor source, the effectiveness of mitigation decreased by 16.9%. Moreover, the decline becomes more pronounced, reaching 25.11% when we overlook community information.

5 CONCLUSION

In this paper, we present an RL framework to eliminate the impact of misinformation on social media platforms, with low user awareness and business cost. We propose a knowledge-informed agent trained with a randomized algorithm, which blocks a minimal set of links to mitigate rumors. Through extensive experiments conducted on real-world social media platforms, our method demonstrates a substantial effectiveness in curbing the spread of misinformation, showcasing an improvement of over 21.1% compared to baseline methods. In the future, we aim to extend our framework to solve more complicated misinformation challenges, such as combating multiple rumor sources in dynamically changing social networks.

ACKNOWLEDGMENTS

This work is supported in part by National Natural Science Foundation of China under grant U23B2030, National Key Research and Development Program of China under grant 2020YFA0711403. This work is also supported in part by Beijing National Research Center for Information Science and Technology (BNRist).

REFERENCES

- [1] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*. 665–674.
- [2] Jingtao Ding, Chang Liu, Yu Zheng, Yunke Zhang, Zihan Yu, Ruikun Li, Hongyi Chen, Jinghua Piao, Huandong Wang, Jiazhen Liu, et al. 2024. Artificial Intelligence for Complex Network: Potential, Methodology and Application. *arXiv preprint arXiv:2402.16887* (2024).
- [3] Jingtao Ding, Yuhuan Quan, Xiangnan He, Yong Li, and Depeng Jin. 2019. Reinforced Negative Sampling for Recommendation with Exposure Data. In *IJCAI*. Macao, 2230–2236.
- [4] Herbert W Hethcote. 2000. The mathematics of infectious diseases. *SIAM review* 42, 4 (2000), 599–653.
- [5] Masahiro Kimura, Kazumi Saito, and Hiroshi Motoda. 2008. Minimizing the spread of contamination by blocking links in a network. In *AAAI*, Vol. 8.
- [6] Jure Leskovec and Julian McAuley. 2012. Learning to discover social circles in ego networks. *Advances in neural information processing systems* 25 (2012).
- [7] Dong Li, Cuihua Wang, Shengping Zhang, Guanglu Zhou, Dianhui Chu, and Chong Wu. 2017. Positive influence maximization in signed social networks based on simulated annealing. *Neurocomputing* 260 (2017), 69–78.
- [8] Ying Liu, Ming Tang, Tao Zhou, and Younghae Do. 2015. Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics. *Scientific reports* 5, 1 (2015), 13172.
- [9] Eli Meir, Hagai Maron, Shie Mannor, and Gal Chechik. 2021. Controlling graph dynamics with reinforcement learning and graph neural networks. In *International Conference on Machine Learning*. PMLR, 7565–7577.
- [10] Lawrence Page. 1998. The pagerank citation ranking: Bringing order to the web. Technical report. *Stanford Digital Library Technologies Project*, 1998 (1998).
- [11] Priyanka Parimi and Rashmi Ranjan Rout. 2021. Genetic algorithm based rumor mitigation in online social networks through counter-rumors: a multi-objective optimization. *Information Processing & Management* 58, 5 (2021), 102669.
- [12] Yuhuan Quan, Jingtao Ding, Chen Gao, Lingling Yi, Depeng Jin, and Yong Li. 2023. Robust Preference-Guided Denoising for Graph based Social Recommendation. In *Proceedings of the ACM Web Conference 2023*. 1097–1108.
- [13] Hanghang Tong, B Aditya Prakash, Tina Eliassi-Rad, Michalis Faloutsos, and Christos Faloutsos. 2012. Gelling, and melting, large graphs by edge manipulation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 245–254.
- [14] Ruidong Yan, Yi Li, Weili Wu, Deying Li, and Yongcai Wang. 2019. Rumor blocking through online link deletion on social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13, 2 (2019), 1–26.
- [15] Zhenhua Yu, Si Lu, Dan Wang, and Zhiwu Li. 2021. Modeling and analysis of rumor propagation in social networks. *Information Sciences* 580 (2021), 857–873.