# Revisiting Fake News Detection: Towards Temporality-aware Evaluation by Leveraging Engagement Earliness

Junghoon Kim*, Junmo Lee*, Yeonjun In, Kanghoon Yoon, Chanyoung Park†

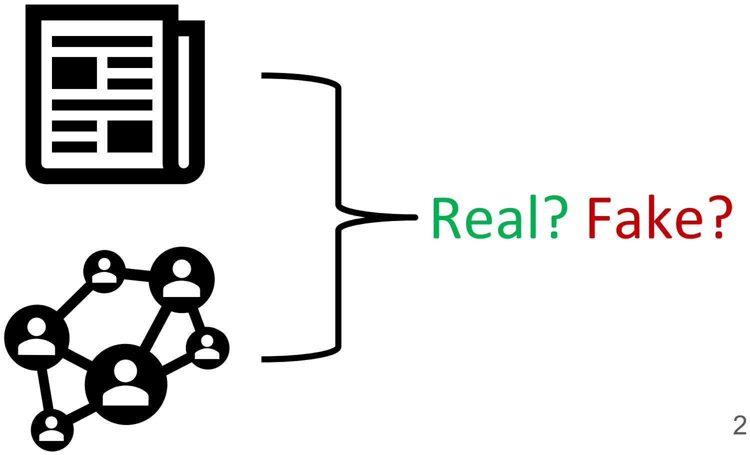
Korea Advanced Institute of Science & Technology (KAIST)

{jhkim611, bubblego0217, yeonjun.in, ykhoon08, cy.park}@kaist.ac.kr

# Introduction


- Fake News Detection
  - Task of **identifying news articles containing false information**
    - Vital for **societal security, public health,** etc.
  - Binary classification of “veracity labels” (0: real, 1: fake)
  - Content-based
    - Leverage **patterns from the news article text itself**
    - E.g., semantic representations, emotional features
  - Social graph-based
    - Additionally **utilize social context knowledge**
    - E.g., user info, retweets

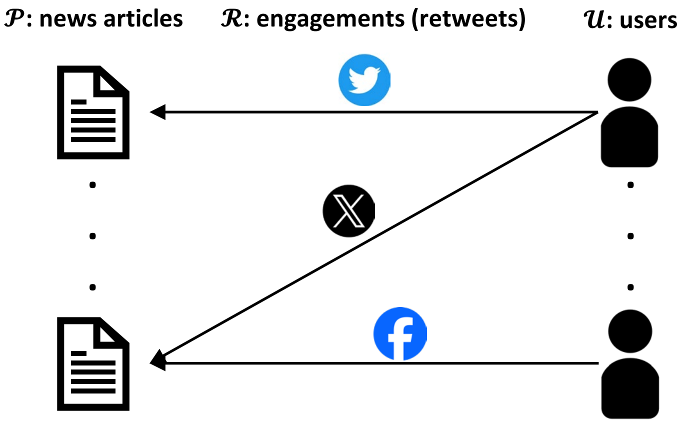
→ **Model such social contexts into graph structures**



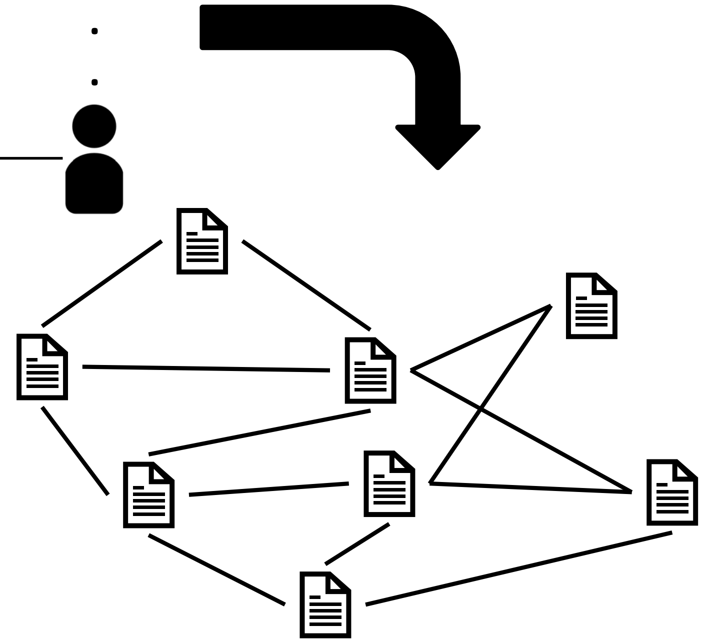

https://www.surrey.ac.uk/news/fighting-fake-news-online-going-beyond-scientific-slogans-connect-people-emotionally

# Introduction

- Fake news dataset $\mathcal{D}$



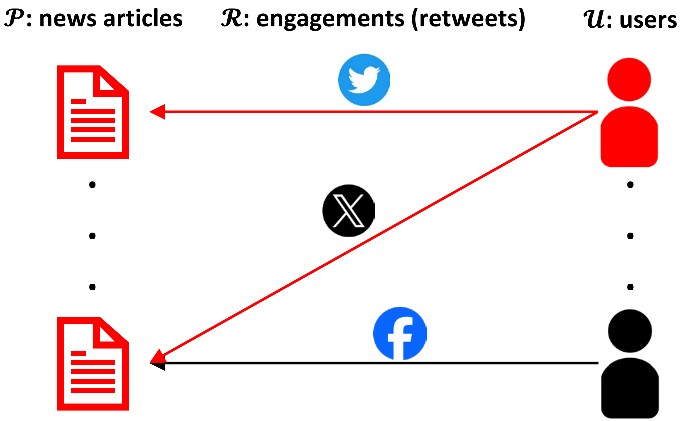

- Social graph $\mathcal{G} = (\mathcal{P}, \mathcal{A})$
    - **Nodes are news articles**



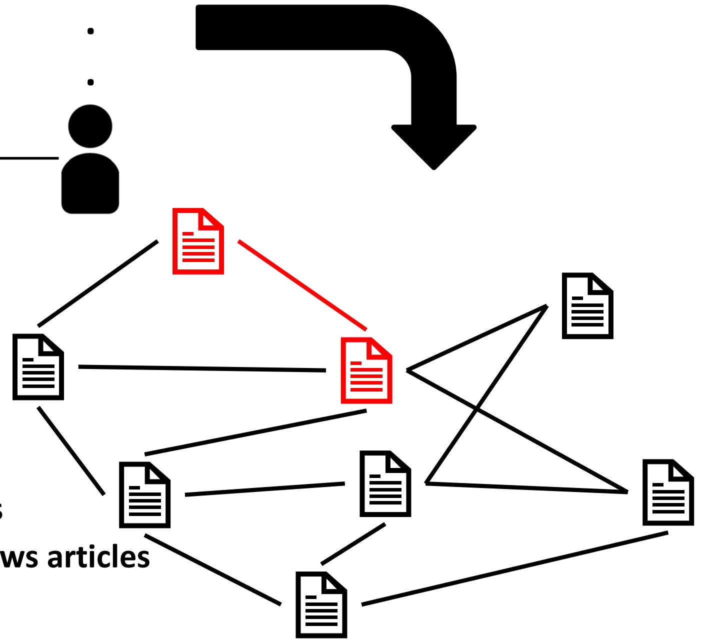
Wu, Jiaying, and Bryan Hooi. "Decor: Degree-corrected social graph refinement for fake news detection." Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023.

# Introduction

- Fake news dataset $\mathcal{D}$




- Social graph $\mathcal{G} = (\mathcal{P}, \mathcal{A})$
  - **Nodes are news articles**
  - Adjacency matrix $\mathcal{A} \in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{P}|}$
    - **Edge weights represent # of co-user engagements (users who retweeted both) between a pair of news articles**

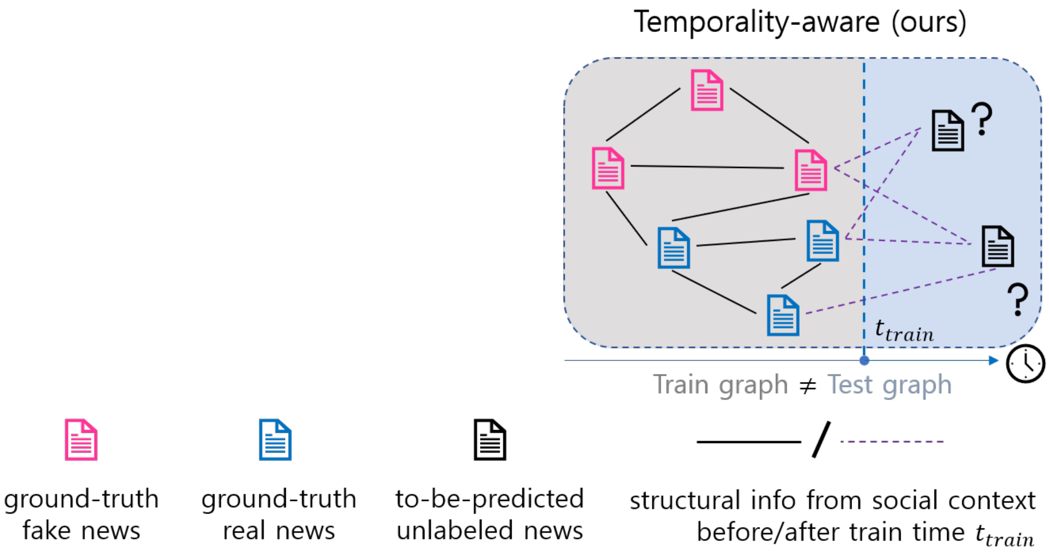




Wu, Jiaying, and Bryan Hooi. "Decor: Degree-corrected social graph refinement for fake news detection." Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023.

# Introduction

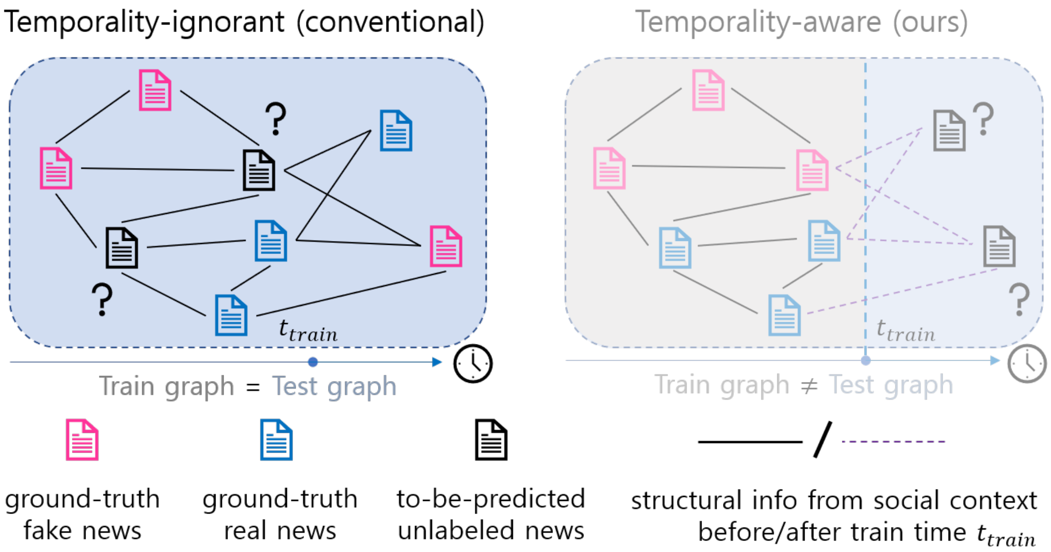
- Shortcomings in previous works
  - Conventional social graph-based methods are **inconsistent with real-world scenarios**
    - In practice, a model would only be trained with data up to a specific point in time (collected in advance), with **future data only available at test time,** i.e., ***temporality-aware*** setting



# Introduction

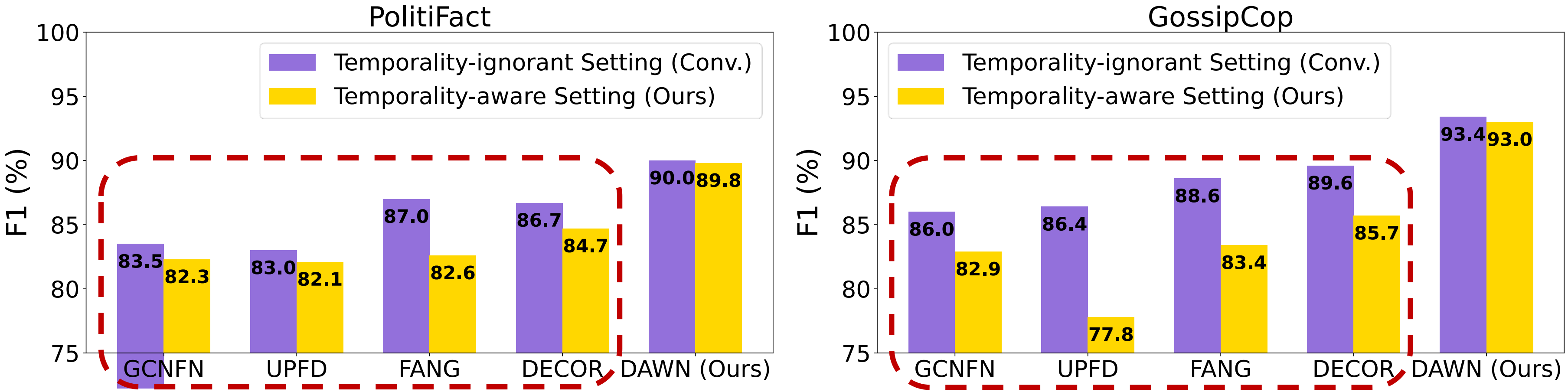
- Shortcomings in previous works
  - Conventional social graph-based methods adopt a ***temporality-ignorant*** setting, assuming **access to future data**
    - Article-related (e.g., textual contents & veracity labels): train / test **divided via random split**
    - Context related (e.g., users, tweets): using **social contexts after training time during training**

**→ Data leakage!**



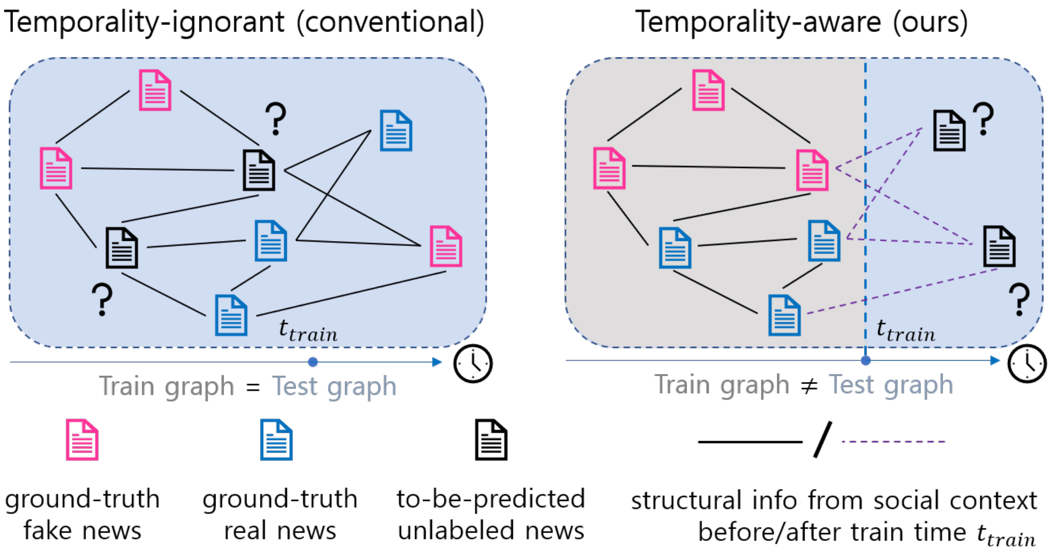
# Introduction

- Shortcomings in previous works
  - Conventional social graph-based methods suffer a **sharp decrease in performance** under this new setting
    - **Why?** Single **fixed structure (patterns) for both training & testing** (inherent design flaw)
      → **Substantial change in structure after training** (more realistic)



# Introduction

- In a nutshell,
  - Conventional social graph-based methods are **unrealistic & unfit** under real-world, temporality-aware settings
    - Where news articles & social context data (+ graph structure) are split by time



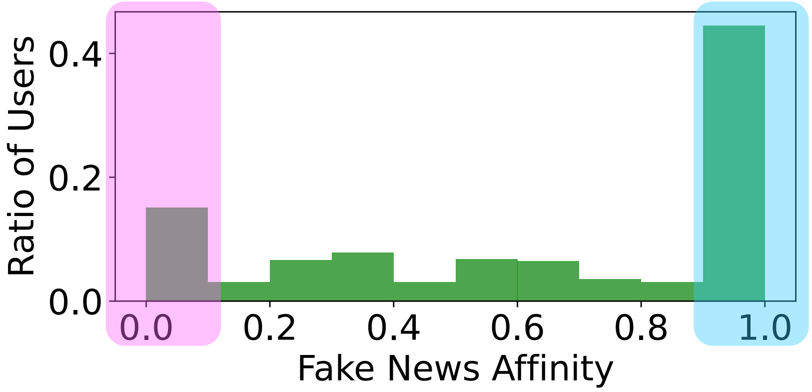
# Introduction

- In a nutshell,
  - Conventional social graph-based methods are **unrealistic & unfit** under real-world, temporality-aware settings
    - Where news articles & social context data (+ graph structure) are **split by time**
    - **Why?** Single **fixed structure (patterns) for both training & testing** (inherent design flaw)
      → **Substantial change in structure after training** (more realistic)
    - Need for **more robust** features & modeling

- Solution: let's **exploit *time-independent* patterns**, i.e., <u>engagement earliness</u>!

# Hypothesis

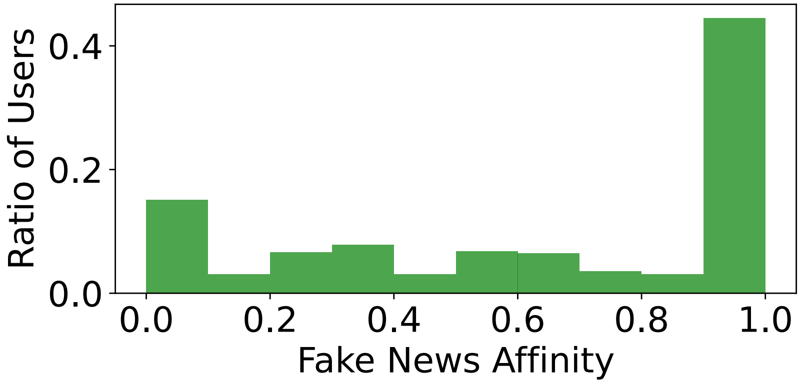
- Fake News Affinity (FNA) score for each active user $u \in \mathcal{U}$
  - $FNA(u) = \frac{\text{\# of engagements with fake news by } u}{\text{\# of all engagements by } u}$ (↑ FNA ≈ ↑ attraction to fake news)




- **Confirmation bias**: people have a **tendency to engage with either** only fake news **or** real news
  - General, well-known social behavior, i.e., **independent of time**

# Hypothesis

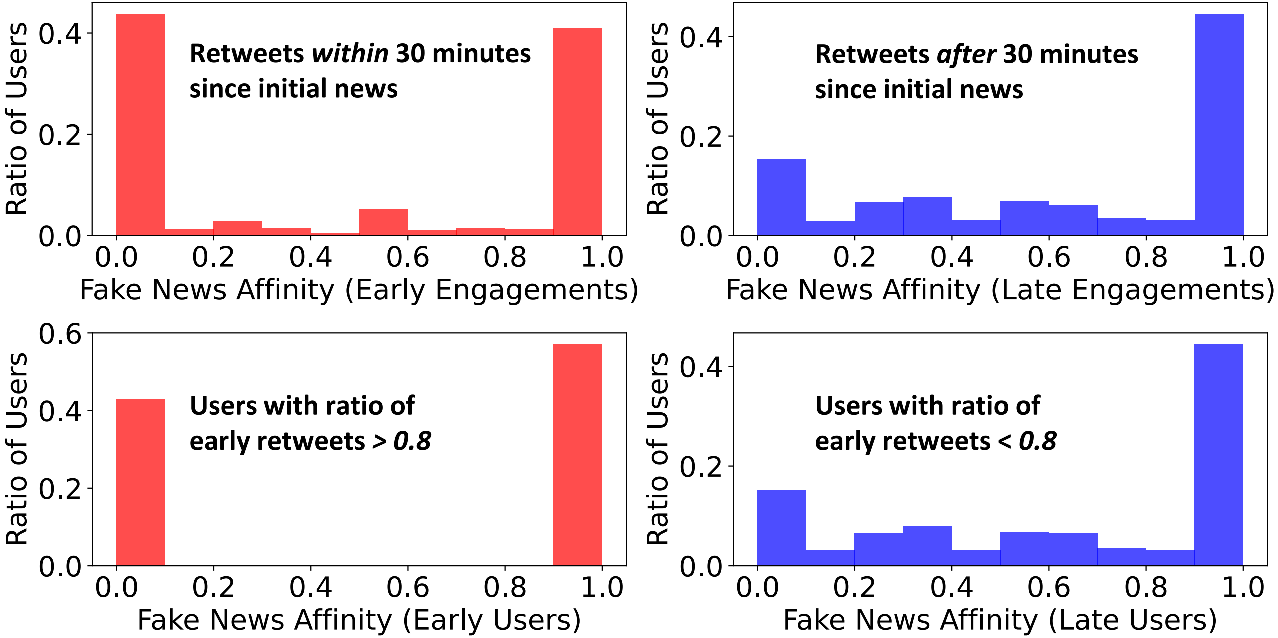
- Fake News Affinity (FNA) score for each active user $u \in \mathcal{U}$
  - $FNA(u) = \frac{\#\text{ of engagements with fake news by } u}{\#\text{ of all engagements by } u}$ (↑ FNA ≈ ↑ attraction to fake news)




- **Hypothesis: Confirmation bias would amplify in users displaying earlier engagement patterns**
  - *Stronger* opinions & beliefs would lead to *quicker* news consumption & responses e.g., retweets
    - More polarized!

# Data Analysis

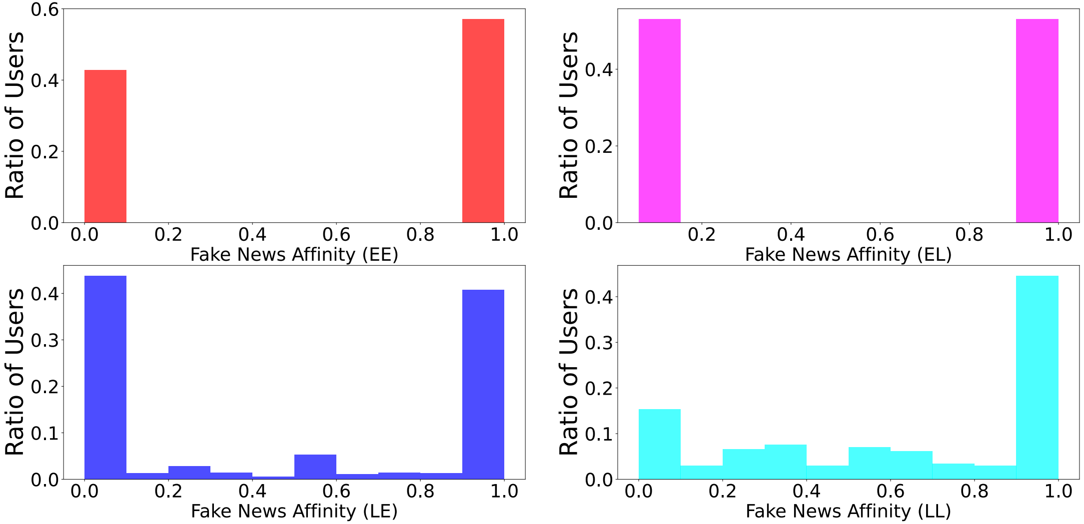
- Two perspectives: **engagement-wise** & **user-wise** earliness patterns
  - Engagement-wise (deadline)
  - User-wise (threshold)




  - The **skewed tendency intensifies with early engagements (users) over late engagements (users)**

# Data Analysis

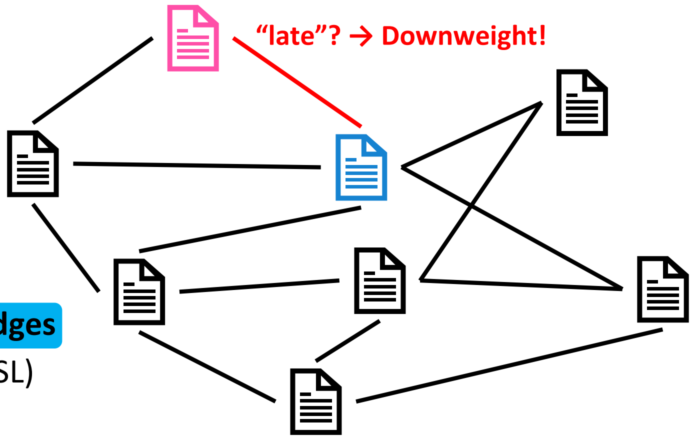
- **Joint** earliness patterns
  - Four groups
    - Early users' early engagements (EE), early users' late engagements (EL)
    - Late users' early engagements (LE), late users' late engagements (LL)




  - **Further capture potentially missed patterns**: EL is more skewed than LL, LE is more skewed than LL

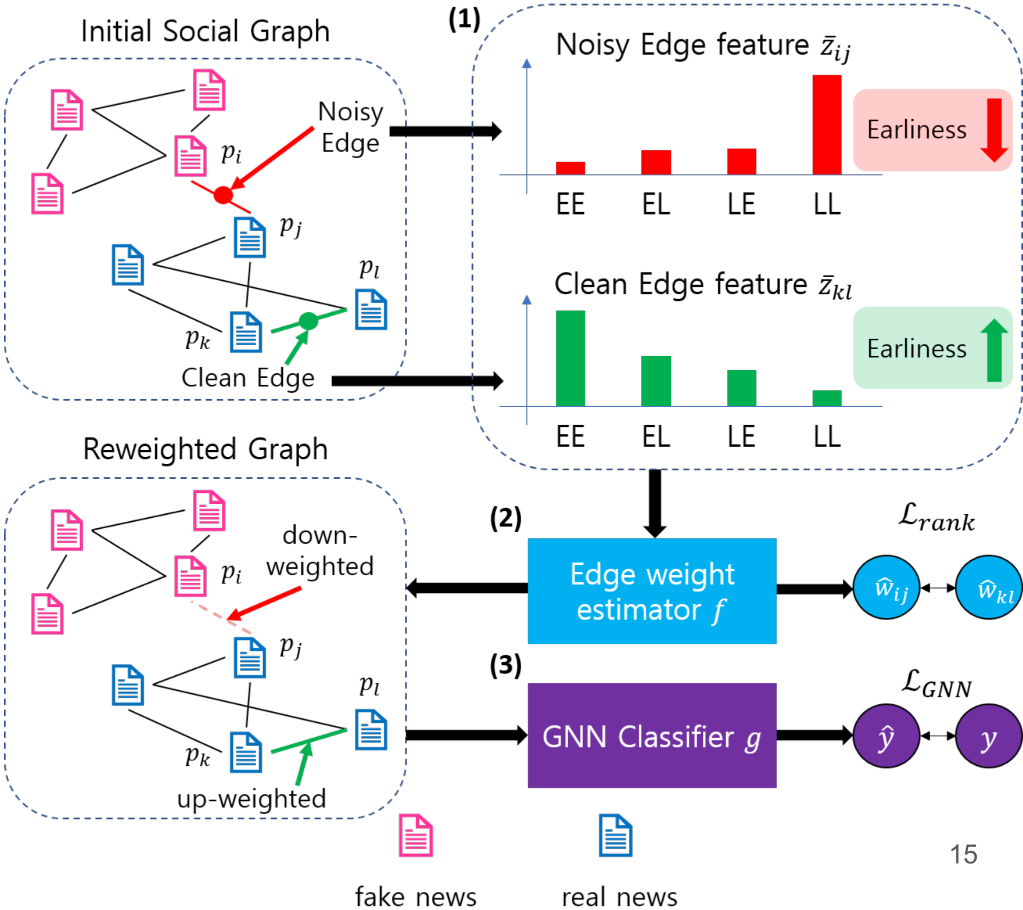
# Data Analysis

- **Implications**
  - *Earlier* user engagements imply a *stronger confirmation bias*, i.e., a **stronger tendency to link news articles of the same veracity** (clean edges) within the social graph
  - Edges containing *later* user engagements have a **higher likelihood of connecting "real" – "fake" news article nodes** (noisy edges)
  - Such underlying patterns represent fundamental social behaviors and thus **occur similarly regardless of time**
  - Utilizing this, let's **identify and adjust the weights of these noisy edges** that hinder detection performance → Graph Structure Learning (GSL)



# Proposed method: DAWN

- **D**etecting fake news via e**A**rliness-guided re**W**eighti**N**g
  - Edge feature construction
  - Reweighting via edge weight estimator $f$
  - Fake news detection via GNN classifier $g$

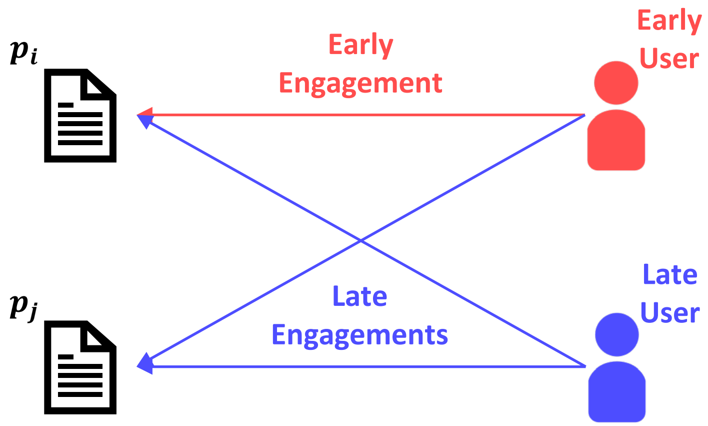


# Proposed method: DAWN

I. Edge feature construction

- Reflects our previous findings regarding engagement earliness
- $\bar{\mathbf{z}}_{ij} \in \mathbb{R}^{\mathbf{4}}$ for existing edge between node pair $p_i, p_j$ obtained by
    - Dividing the engagements into four groups (EE, EL, LE, LL)
    - Concatenating the size of each group
      + column-wise normalization

→ **Represents the earliness profile of each edge**




[EE, EL, LE, LL] = [1, 1, 0, 2]

↓

$\bar{z}_{ij} \in \mathbb{R}^4$

# Proposed method: DAWN

II. Noisy edge suppression

- Obtain adjusted edge weights: $w_{ij} = f\left(\bar{z}_{ij}\right) = sigmoid(MLP(\bar{z}_{ij}))$
  - $w_{ij} \in [0,1]$

- Regularize $f$ via **ranking loss**: $\mathcal{L}_{rank} = \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\max(0, -\left(w_{clean}^{(i)} - w_{noisy}^{(j)}\right) + margin)$
  - Maximize the distance between the weights of sampled clean edges (↑) and noisy edges (↓)
  - Observation of only $K^2$ pairs is sufficient in practice & significantly reduces training time

- Replace the original adjacency matrix with the adjusted weights

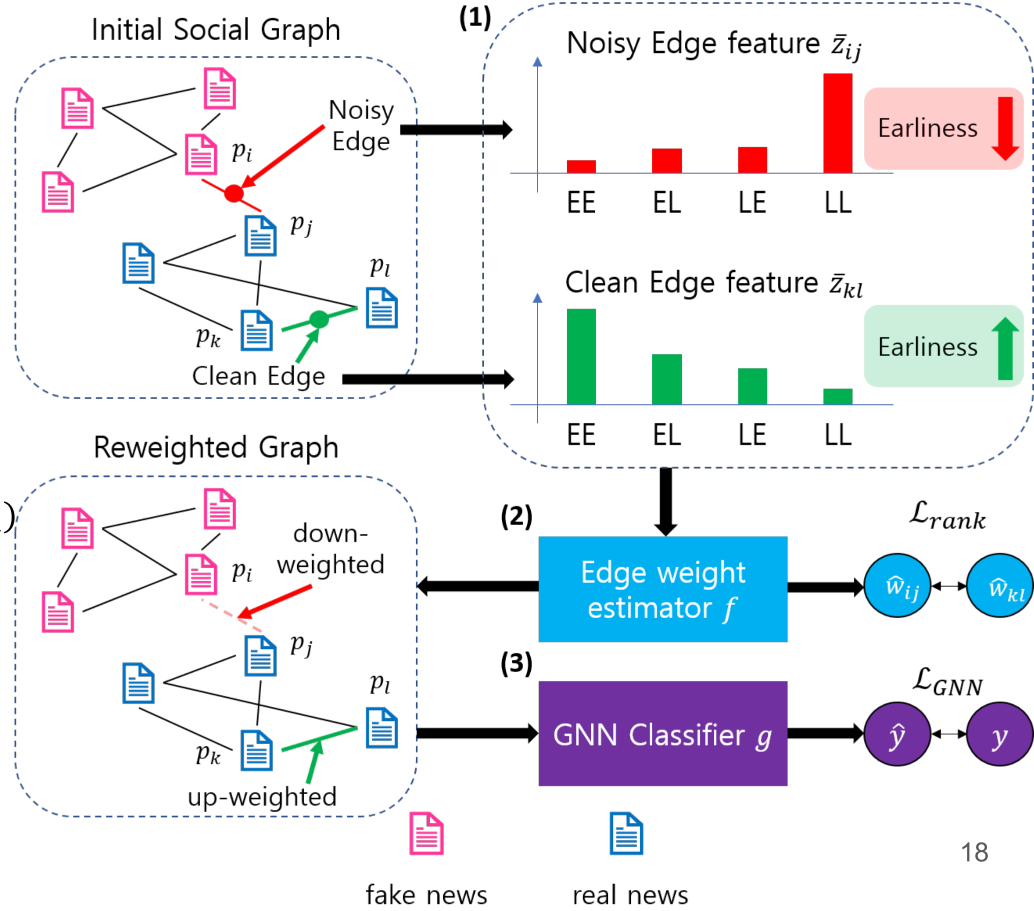
# Proposed method: DAWN

III. Fake news detection

- Initial node features:
  obtained from text via pre-trained BERT

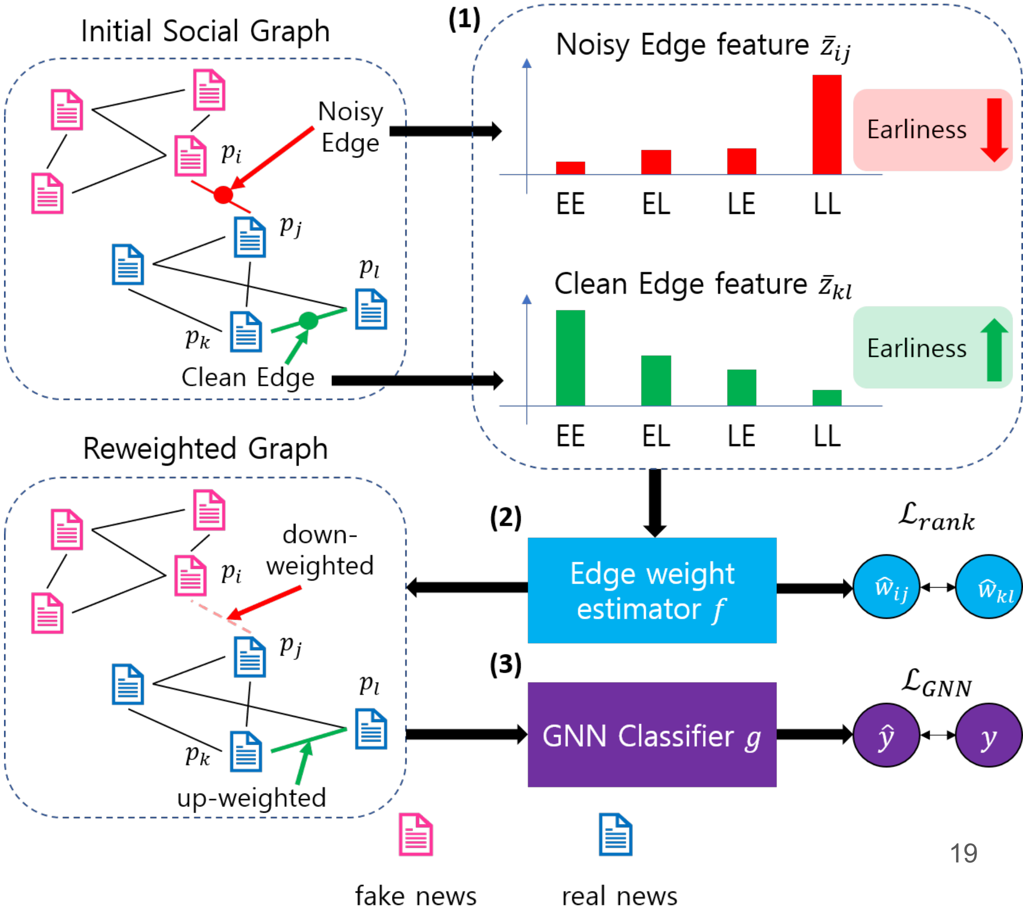
- Inputs for GNN classifier $g$ (e.g., GCN)
  - Training node features
  - Adjusted training adjacency matrix

- **Prediction loss**: $\mathcal{L}_{GNN} = \sum_{p_n \in \mathcal{P}_{train}} l(\hat{y}_n, y_n)$
  - $l(\hat{y}_n, y_n)$: cross entropy loss



# Proposed method: DAWN

- Summary of DAWN
  - $\mathcal{L}_{final} = arg \min_{\theta,\phi} \mathcal{L}_{GNN} + \alpha \mathcal{L}_{rank}$
    - $\alpha$: balancing hyperparameter
  - After splitting the data by time & training on the training graph,
  - The best performing model on the validation graph is used for final prediction on the test graph



# Experiments

- Baselines
  - **G1**: four content-based methods
    - dEFEND\c, DualEmo/c, BERT, GPT3.5
  - **G2**: six social graph-based methods
    - GCNFN, UPFD, FANG, GCN, GAT, GraphSAGE
  - **G3**: two GSL-based methods
    - RS-GNN, DECOR

| Dataset | PolitiFact | GossipCop |
|---|---|---|
| # News Articles | 597 | 8,763 |
| # Real News | 282 | 6,764 |
| # Fake News | 315 | 1,999 |
| # Users | 162,262 | 129,820 |
| # Tweets (Engagements) | 255,227 | 516,172 |

- Datasets
  - **PolitiFact** & **GossipCop**: labeled news articles & related tweets by users on X (Twitter)
  - Temporality-aware evaluation
    - Train : val : test = 70% : 10% : 20% **in temporal order**
    - News articles, users & tweets split accordingly

# Experiments

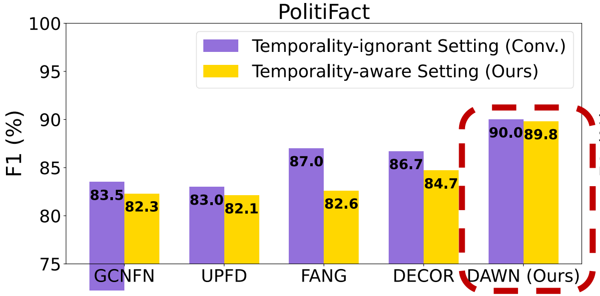
- Detection performance
  - Under temporality-aware settings,
    - Up to 5.6%p (acc.) 7.3%p (F1.) enhancement
    - Robust performance due to **time-independent earliness patterns**

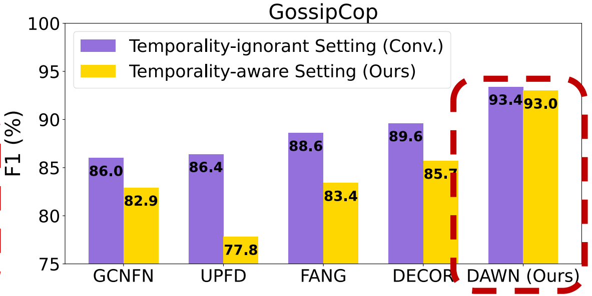
| | Method | PolitiFact | | GossipCop | |
|---|---|---|---|---|---|
| | | acc. | f1. | acc. | f1. |
| G1 | dEFEND\c [24] | 81.2 | 79.6 | 75.4 | 68.3 |
| | DualEmo\c [33] | 84.0 | 81.5 | 77.9 | 71.9 |
| | BERT [6] | 84.2 | 81.7 | 76.4 | 69.1 |
| | GPT3.5 | 75.7 | 77.9 | 62.8 | 56.6 |
| G2 | GCNFN [20] | 84.8 | 82.3 | 85.3 | 82.9 |
| | UPFD [7] | 84.6 | 82.1 | 82.0 | 77.8 |
| | FANG [21] | 85.5 | 82.6 | 86.1 | 83.4 |
| | GCN [15] | 86.4 | 83.0 | 87.4 | 84.7 |
| | GAT [27] | 86.7 | 79.5 | 86.5 | 83.6 |
| | GraphSAGE [11] | 85.4 | 80.9 | 87.5 | 85.5 |
| G3 | RS-GNN [4] | 80.8 | 62.8 | 85.9 | 83.6 |
| | DECOR [31] | 87.4 | 84.7 | 88.1 | 85.7 |
| **Ours** | DAWN | **91.9** | **89.8** | **93.7** | **93.0** |

# Experiments

- Detection performance
  - Under temporality-aware settings,
    - Up to 5.6%p (acc.) 7.3%p (F1.) enhancement
    - Robust performance due to **time-independent earliness patterns**
  - Under temporality-ignorant settings,
    - Significantly outperforms baselines
    - **Marginal performance drop** from previous setting
  - **Versatility of DAWN** under various scenarios

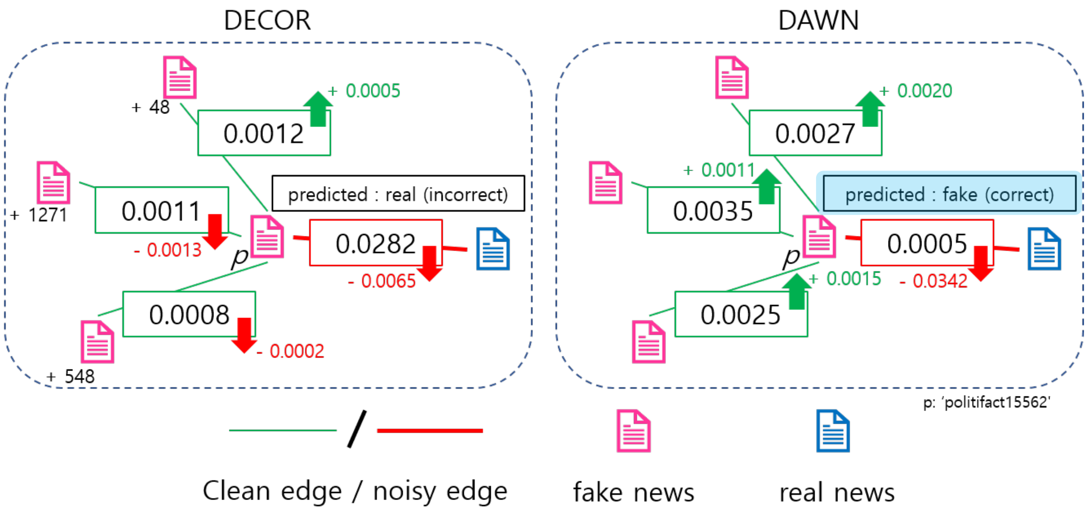
| | Method | PolitiFact | | GossipCop | |
|---|---|---|---|---|---|
| | | acc. | f1. | acc. | f1. |
| **G1** | dEFEND\c [24] | 81.2 | 79.6 | 75.4 | 68.3 |
| | DualEmo\c [33] | 84.0 | 81.5 | 77.9 | 71.9 |
| | BERT [6] | 84.2 | 81.7 | 76.4 | 69.1 |
| | GPT3.5 | 75.7 | 77.9 | 62.8 | 56.6 |
| **G2** | GCNFN [20] | 84.8 | 82.3 | 85.3 | 82.9 |
| | UPFD [7] | 84.6 | 82.1 | 82.0 | 77.8 |
| | FANG [21] | 85.5 | 82.6 | 86.1 | 83.4 |
| | GCN [15] | 86.4 | 83.0 | 87.4 | 84.7 |
| | GAT [27] | 86.7 | 79.5 | 86.5 | 83.6 |
| | GraphSAGE [11] | 85.4 | 80.9 | 87.5 | 85.5 |
| **G3** | RS-GNN [4] | 80.8 | 62.8 | 85.9 | 83.6 |
| | DECOR [31] | 87.4 | 84.7 | 88.1 | 85.7 |
| **Ours** | DAWN | **91.9** | **89.8** | **93.7** | **93.0** |



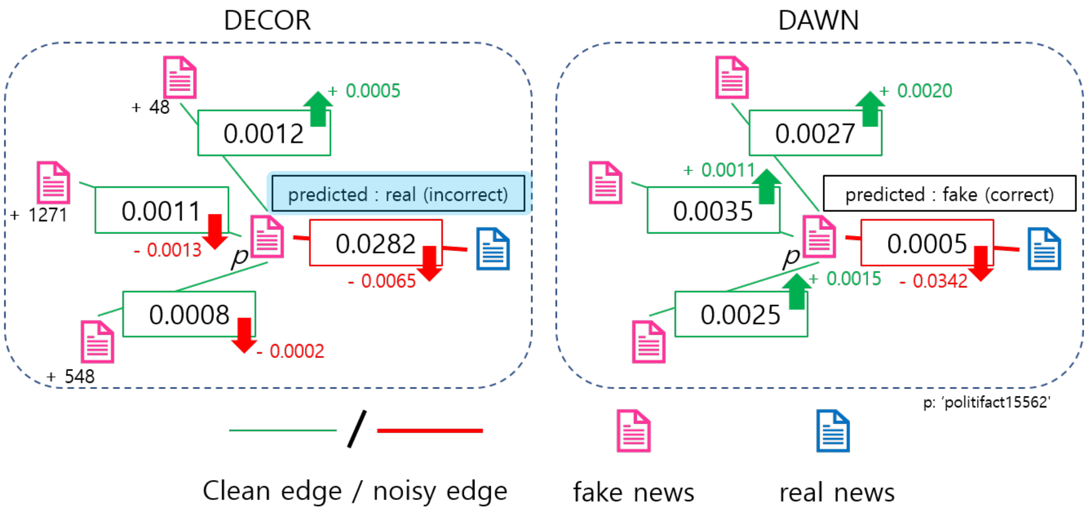
# Experiments

- Case study
  - DAWN's **time-independent** earliness features **successfully identify & reweight edges**, leading to a **correct veracity prediction**



# Experiments

- Case study
  - DAWN's **time-independent** earliness features **successfully identify & reweight edges**, leading to a **correct veracity prediction**
  - In comparison, previous methods wrongly reweight edges due to **substantial changes in graph structure**, leading to **incorrect predictions**



# Conclusion

- We explore the need for a **temporality-aware** setting that better reflects real-world fake news detection scenarios – in which **previous methods suffer significant performance drop**
    - Future data (both article-wise & context wise) should be unavailable during training

- In-depth analyses reveal **time-independent patterns regarding engagement earliness**
    - Rooted in general social behaviors, e.g., confirmation bias
    - Later user engagements contribute more to noisy edges within the social graph

- A novel framework **DAWN** is proposed to successfully utilize such patterns in the form of Graph Structure Learning

- We verify the **versatility & robust effectiveness** of DAWN through extensive experiments

# Thank you for your attention!

Paper

Code

Contact: jhkim611@kaist.ac.kr

# Appendix

- Algorithm

**Algorithm 1** Training Algorithm of DAWN

**Input:** $\mathcal{G}_{train} = (\mathcal{P}_{train}, \mathcal{A}_{train}), \mathcal{G}_{val} = (\mathcal{P}_{val}, \mathcal{A}_{val}), \mathcal{X}$, normalized edge features, $K, \alpha$

**Output:** Edge weight estimator $f$, GNN classifier $g$

```
Randomly initialize the parameters of f and g
for i = 1 to # epochs do
    Get the reweighted training adjacency matrix W_train with f by Equation 3
    Input W_train and training node features from X to g to get veracity predictions
    Randomly sample K clean and noisy edges, respectively
    Jointly optimize parameters θ and ϕ by Equation 6
    Get the reweighted validation adjacency matrix W_val and perform validation on G_val
end for
Return f and g best performing on G_val
```

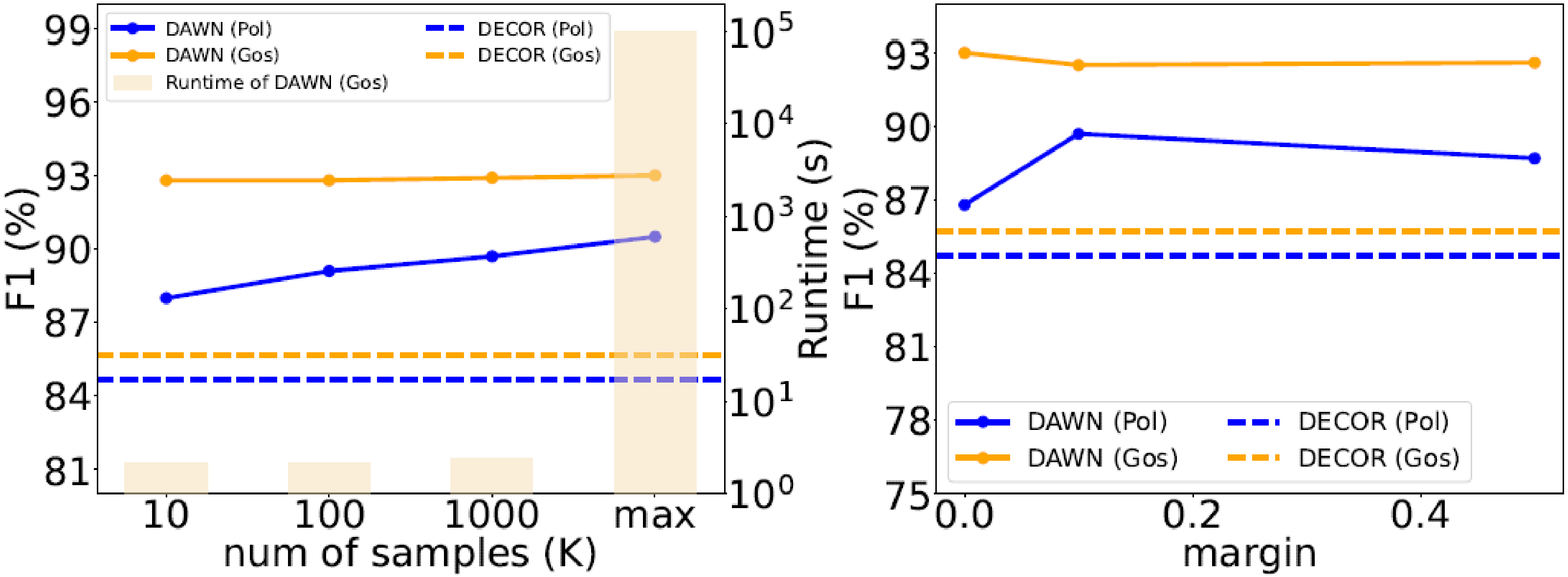
# Appendix

- Ablation study
  - Feature variants
    - **DAWN+RAND**: random values from uniform distribution as edge features
    - **DAWN-USER** / **DAWN-ENG**: remove user- / engagement-wise earliness patterns
    - **DAWN+RATIO**: row-wise column normalization
    - **DAWN+NF**: concatenate node features as edge features
  - Component variants
    - **DAWN-RANK**: remove $\mathcal{L}_{rank}$, i.e., $\alpha = 0$
    - **DAWN+BC**: replace $\mathcal{L}_{rank}$ with binary classification loss

| | | PolitiFact | GossipCop |
|---|---|---|---|
| **Ours** | DAWN | **89.8** | **93.0** |
| **Feature Variants** | DAWN+RAND | 58.9 | 79.7 |
| | DAWN-USER | 85.6 | 92.5 |
| | DAWN-ENG | 83.2 | 92.1 |
| | DAWN+RATIO | 76.1 | 89.2 |
| | DAWN+NF | 72.0 | 80.7 |
| **Component Variants** | DAWN-RANK | 76.5 | 91.9 |
| | DAWN+BC | 62.4 | 90.2 |

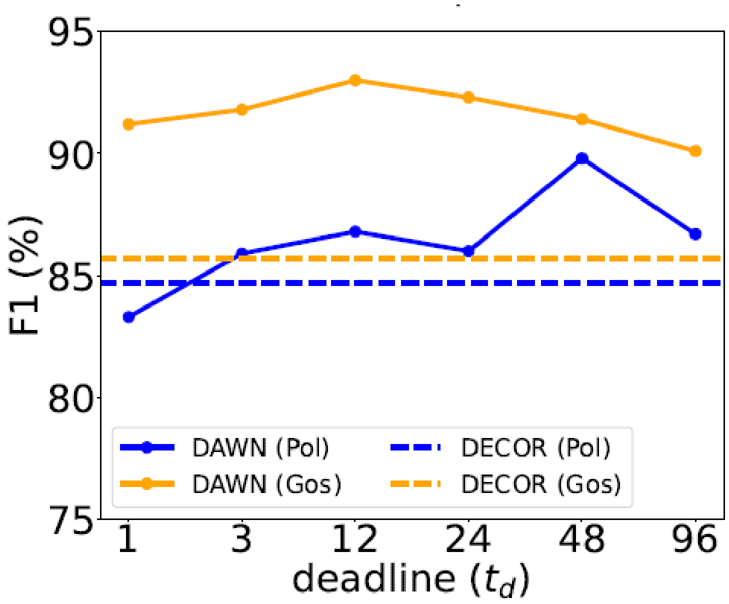
# Appendix

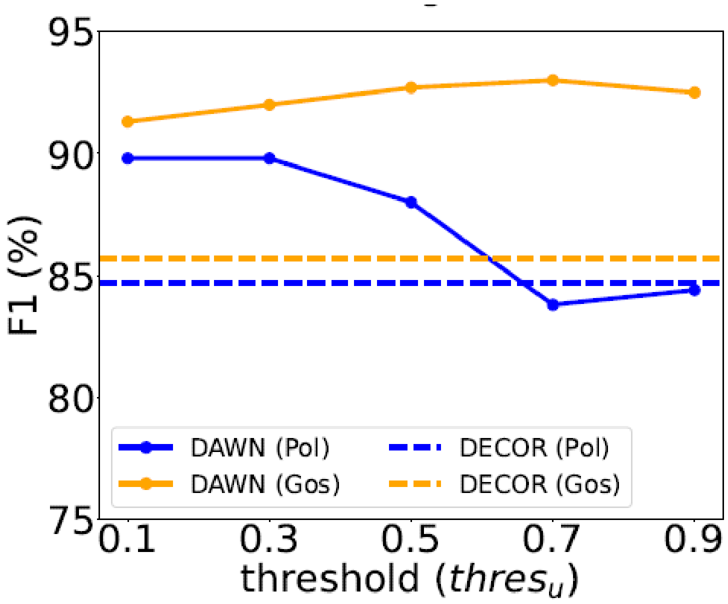
- Hyperparameter sensitivity ($K, margin$)



# Appendix

- Hyperparameter sensitivity ($\mathrm{t_d}, thres_u$)

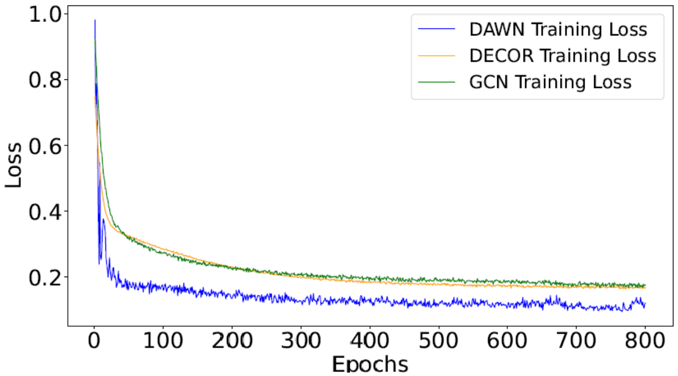


# Appendix

- Efficiency on large networks

| Method | GossipCop | |
|---|---|---|
| | Runtime (s) | f1. |
| GCN | 0.286 | 84.7 |
| DECOR | 0.544 | 85.7 |
| DAWN | 2.41 | 93.0 |



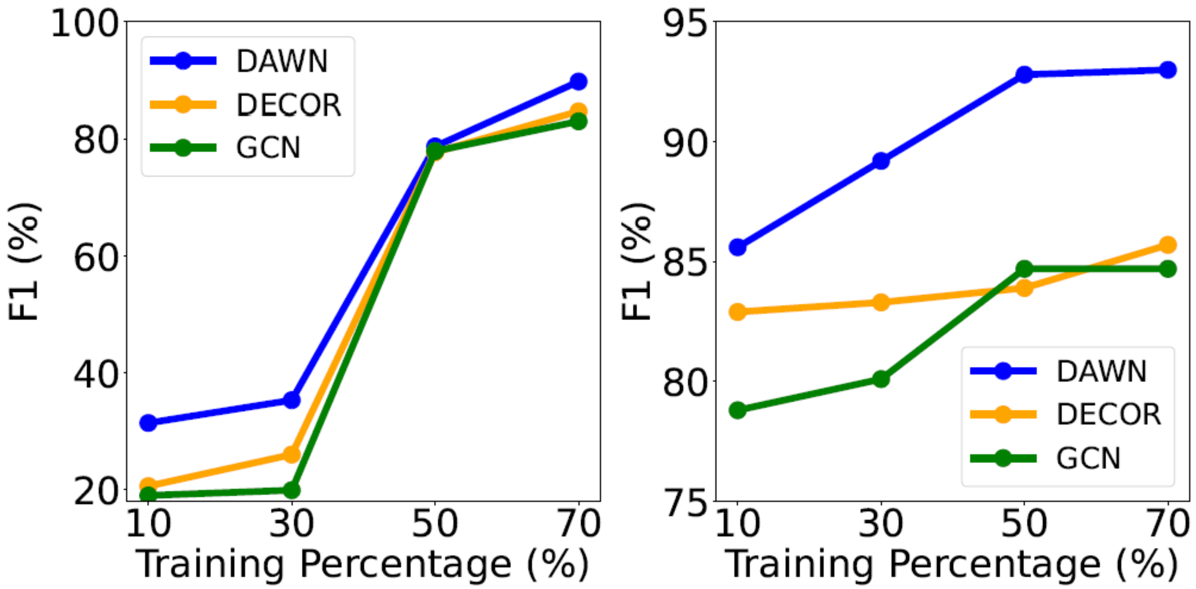
# Appendix

- Homophily comparison after adjustment

| | Original graph | Adjusted by DECOR | Adjusted by DAWN |
|---|---|---|---|
| PolitiFact | 0.724 | 0.807 | 0.821 |
| GossipCop | 0.932 | 0.943 | 0.954 |

# Appendix

- Performance under limited training data



# Appendix

- Generalizability (Chinese dataset MCFEND)

| Dataset | MCFEND |
|---|---|
| # News Articles | 23,789 |
| # Real News | 6,074 |
| # Fake News | 17,715 |
| # Users | 803,779 |
| # Engagements | 2,102,902 |