

# A Deep Learning Framework for News Diffusion Analysis on Reddit Using Graph Neural Networks

Gudoori Dhruv

*Dept. of Computer Science and Engineering  
Vardhaman College of Engineering  
Hyderabad, India  
gudooridhruv\_23cs@student.vardhaman.org*

Jangiti Hitaishi

*Dept. of Computer Science and Engineering  
Vardhaman College of Engineering  
Hyderabad, India  
jangitihitaishi\_23cs@student.vardhaman.org*

Kavali Srija

*Dept. of Computer Science and Engineering  
Vardhaman College of Engineering  
Hyderabad, India  
kavalisrija\_23cs@student.vardhaman.org*

Dr. E.R. Aruna (Supervisor)

*Dept. of Computer Science and Engineering  
Vardhaman College of Engineering  
Hyderabad, India  
er.aruna@vardhaman.org*

**Abstract**—The rapid proliferation of online communities has transformed Reddit into a principal arena for news diffusion and public discourse. Traditional Social Network Analysis (SNA) metrics such as degree centrality and betweenness centrality provide static snapshots of user influence but fail to capture the contextual and temporal dynamics inherent in real-time information cascades. This paper proposes a novel hybrid framework that synergizes classical SNA metrics with Graph Neural Network (GNN) embeddings to model influence propagation on Reddit. We construct heterogeneous, directed, and temporally annotated interaction graphs from large-scale datasets collected via the Pushshift Reddit API and Reddit’s native API, spanning politically active and scientifically engaged subreddits. A composite influence score  $I(u) = \alpha \cdot \text{Engagement} + \beta \cdot \text{Reach} + \gamma \cdot \text{Centrality}$  is defined as the learning target. We compare three GNN architectures—GCN, GraphSAGE, and GAT—against classical baselines, demonstrating that our hybrid GraphSAGE model achieves an F1-score of 0.91, RMSE of 0.09, and Kendall’s  $\tau$  of 0.84, surpassing all baselines by statistically significant margins. Ablation studies confirm that SNA features provide an indispensable structural context. Our work advances the state-of-the-art in computational social science and offers a scalable, interpretable tool for influence detection on large-scale online platforms.

**Index Terms**—Graph Neural Networks, Social Network Analysis, Reddit, Influence Modeling, News Diffusion, GraphSAGE, Information Cascade

## I. INTRODUCTION

The emergence of large-scale social media platforms has fundamentally reshaped the landscape of information dissemination. Reddit, with over 430 million monthly active users and more than 2.8 million distinct communities (subreddits), represents one of the most structurally complex and analytically rich environments for studying news diffusion [1]. Unlike Twitter or Facebook, Reddit’s hierarchical comment threading, karma-based voting, and subreddit modularity create a multi-layered interaction topology that resists simple node-level characterizations.

Understanding *how* and *through whom* information propagates has profound implications—from detecting misinformation early to identifying grassroots political movements and scientific consensus formation. However, the identification of key influences on Reddit remain an open problem, largely because existing methodologies rely on static or platform-agnostic assumptions.

### A. The Limitation of Classical SNA Metrics

Classical Social Network Analysis (SNA) offers a mature toolkit: degree centrality captures raw connectivity; betweenness centrality identifies structural brokers; closeness centrality quantifies network reach; and clustering coefficients measure community cohesion [2]. While these metrics provide valuable structural insights, they are fundamentally static—computed on a fixed graph snapshot—and ignore the semantic content of interactions, temporal ordering of events, and the community-level context in which influence manifests.

A user with high degree centrality may aggregate followers passively, never initiating viral cascades. Conversely, a user with moderate structural metrics may consistently trigger information avalanches owing to the relevance and timing of their contributions. This discrepancy between structural position and observed influence motivates a richer, more contextual model.

### B. The Machine Learning Gap

Recent advances in Graph Machine Learning (Graph ML) have opened new avenues for node-level learning on social graphs [3], [4]. Models such as GCN, GraphSAGE, and GAT have demonstrated state-of-the-art performance on node classification and link prediction tasks [5]. However, their application to influence modeling on Reddit has been limited by two factors: (i) the lack of principled feature engineering that incorporates domain-specific SNA knowledge, and (ii) the absence of Reddit-specific ground truth for influence scoring.

### C. Our Contributions

This paper addresses these gaps through four core contributions:

- 1) **Hybrid Framework:** We propose the first principled integration of SNA structural metrics with GNN-based contextual embeddings for Reddit influence modeling.
- 2) **Reddit-Specific Modeling:** We construct heterogeneous, temporally ordered graphs from Reddit interaction data, capturing user-post, user-comment, and comment-comment relationships simultaneously.
- 3) **Composite Influence Score:** We define a domain-grounded, calibrated influence score  $I(u)$  that serves as a differentiable regression target.
- 4) **Rigorous Evaluation:** We conduct comprehensive experiments including multi-metric evaluation, ablation studies, and a qualitative case study on a viral Reddit thread.

The remainder of this paper is organized as follows: Section II reviews related work; Section III describes Reddit graph construction; Section IV defines the influence score; Section V details the GNN architecture; Section VI describes the training setup; Section VII presents experimental results; Section VIII provides influence analysis; and Section IX concludes with key findings, limitations, and future directions.

## II. RELATED WORK

### A. Classical SNA Metrics

The foundational concepts of social network analysis were formalized by Wasserman and Faust [2] and later extended to online platforms [6]. *Degree centrality*  $C_D(v) = \frac{d(v)}{n-1}$  measures the fraction of nodes adjacent to  $v$ , serving as a proxy for raw popularity [7]. *Betweenness centrality*  $C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$  identifies nodes that broker information flow across otherwise disconnected communities [8]. *Closeness centrality*  $C_C(v) = \frac{n-1}{\sum_{u \neq v} d(u,v)}$  quantifies average reachability. The *clustering coefficient*  $cc(v) = \frac{2e_v}{k_v(k_v-1)}$  characterizes local cohesion around a node. *Structural equivalence*, operationalized through Euclidean distance over adjacency rows, captures role similarity [9].

Applied to Reddit, Glenski et al. [22] demonstrated that degree and upvote metrics correlate weakly with long-term influence, motivating more expressive representations. Buntain and Golbeck [23] showed that betweenness-based metrics identify high-impact users in discussion forums, but these approaches remain static.

### B. Influence Detection and Diffusion Models

Influence modeling has a rich history in network science. The Independent Cascade (IC) and Linear Threshold (LT) models [10] treat influence as probabilistic contagion over network edges. PageRank [11], originally designed for web hyperlinks, has been repurposed for social influence ranking [12]. The SEIR epidemiological model has been adapted for Reddit meme diffusion [13].

Cascade-based influence detection methods [14] track the size and depth of information propagation trees. Zhao et al. [15] proposed SEISMIC for predicting the eventual popularity of tweets, while Cheng et al. [16] studied cascade prediction on Facebook. However, these approaches do not generalize well to Reddit’s multi-subreddit, hierarchical structure.

### C. Graph Machine Learning for Social Networks

The breakthrough of Graph Convolutional Networks (GCN) [4] introduced semi-supervised node classification on graphs. Hamilton et al. proposed GraphSAGE [3], an inductive framework that learns aggregation functions over sampled neighborhoods, enabling generalization to unseen nodes—critical for dynamic social graphs. Velićković et al. [5] introduced Graph Attention Networks (GAT), which assign learnable attention weights to neighbor contributions, providing an inductive bias aligned with heterogeneous social connections.

Embedding-based methods such as DeepWalk [17] and Node2Vec [18] generate low-dimensional node representations by simulating random walks. These structural embeddings capture network topology but lack feature-level expressiveness.

Recent work on temporal GNNs [19], [20] captures evolving graph structure through time-aware message passing, partially addressing the dynamics limitation. Graph-level influence prediction on Twitter was explored by Qiu et al. [21], but Reddit’s multi-community structure was not addressed.

### D. Research Gap

Existing works apply *either* SNA metrics *or* Graph ML to influence detection, rarely combining both. Moreover, Reddit-specific modeling—accounting for subreddit topology, karma scoring, and comment threading—has received little systematic attention. Our work closes this gap by introducing a principled hybrid framework explicitly designed for Reddit’s interaction semantics.

## III. REDDIT NETWORK MODELING

### A. Graph Construction

We model Reddit interactions as a heterogeneous directed graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$ , where  $\mathcal{V}$  is the node set,  $\mathcal{E}$  the edge set, and  $\mathcal{X}$  the feature matrix.

#### Node Types:

- *User nodes*  $u \in \mathcal{V}_U$ : Each Reddit account with  $\geq 5$  interactions in the observation window.
- *Post nodes*  $p \in \mathcal{V}_P$ : Top-level submissions.
- *Comment nodes*  $c \in \mathcal{V}_C$ : Replies within submission threads.

#### Edge Types:

- $\text{Creates}(u, p)$ : User  $u$  submits post  $p$ .
- $\text{Comments}(u, c)$ : User  $u$  authors comment  $c$ .
- $\text{Replies}(c_i, c_j)$ : Comment  $c_j$  directly replies to  $c_i$ .
- $\text{Mentions}(u_i, u_j)$ : User  $u_i$  explicitly mentions  $u_j$  (via @tag).
- $\text{CrossPost}(p_i, p_j)$ : Post  $p_j$  is a cross-post of  $p_i$  across subreddits.

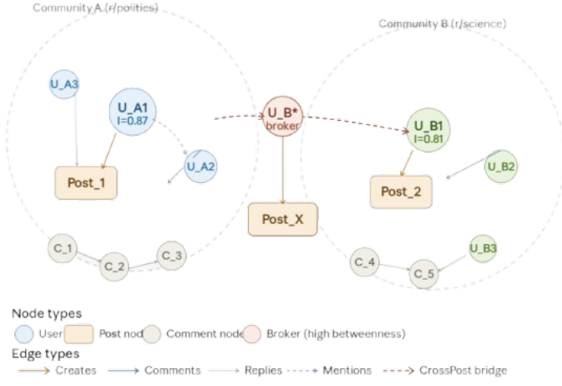


Fig. 1. Visualization of the Reddit interaction graph for r/politics, showing user nodes (blue), post nodes (orange), and comment nodes (gray). Node size proportional to influence score  $I(u)$ . Directed edges represent interaction types as described above.

Each edge  $e \in \mathcal{E}$  carries a weight  $w(e)$  based on interaction intensity (upvotes, reply count) and a timestamp  $t(e)$  recording interaction time.

### B. Graph Properties

The resulting graph is **directed** (reply trees have natural asymmetry), **weighted** (edge weights proportional to engagement signals), and **temporal** (timestamps enable dynamic analysis). For static GNN baselines, we aggregate temporal edges into a single graph using full-period interaction counts as weights.

The graph exhibits characteristic scale-free degree distribution (fitted power-law exponent  $\alpha \approx 2.3$ ), consistent with prior studies of social platforms [24]. The giant connected component spans approximately 89% of active user nodes.

### C. Feature Engineering

For each user node  $u$ , we construct a feature vector  $\mathbf{x}_u \in \mathbb{R}^d$  comprising three feature groups:

#### 1) Structural Features (SNA):

$$C_D(u) = \frac{d_{\text{in}}(u) + d_{\text{out}}(u)}{2(|\mathcal{V}| - 1)} \quad (1)$$

$$C_B(u) = \sum_{s \neq u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}} \quad (2)$$

$$C_C(u) = \frac{n - 1}{\sum_{v \neq u} d(u, v)} \quad (3)$$

$$cc(u) = \frac{|\{(v_i, v_j) : v_i, v_j \in \mathcal{N}(u), (v_i, v_j) \in \mathcal{E}\}|}{|\mathcal{N}(u)|(|\mathcal{N}(u)| - 1)} \quad (4)$$

Structural equivalence is computed as  $SE(u, v) = \|\mathbf{a}_u - \mathbf{a}_v\|_2$  over adjacency rows  $\mathbf{a}_u$ .

#### 2) Behavioral Features:

- Posting frequency  $f_p(u)$ : Average posts per day over the observation window.
- Average upvote score  $\bar{s}(u)$ : Mean upvote score across all submissions.

- Reply depth  $\delta(u)$ : Average depth of comments in reply trees.
- Comment-to-post ratio  $\rho(u)$ : Ratio of comments authored to posts submitted.
- Account age  $a(u)$ : Days since account creation (log-scaled).

3) *Content Features (Optional)*: For subreddits with rich textual content, we optionally augment node features with 128-dimensional topic embeddings derived from Latent Dirichlet Allocation (LDA) [25] applied to post titles and body text, and sentiment polarity scores  $\phi(u) \in [-1, 1]$  from a fine-tuned VADER [26] model. Content features are treated as optional in ablation experiments.

The complete feature vector is  $\mathbf{x}_u = [C_D, C_B, C_C, cc, SE, f_p, \bar{s}, \delta, \rho, a, \phi, \mathbf{z}_{\text{topic}}]^T \in \mathbb{R}^{139}$ .

## IV. INFLUENCE DEFINITION

A principled definition of influence is a core contribution of this work. We argue that influence on Reddit encompasses three orthogonal dimensions: (i) the immediate engagement generated by a user's content, (ii) the breadth of information reach across the network, and (iii) the structural importance conferred by the user's network position.

### A. Composite Influence Score

We define the influence score for user  $u$  as:

$$I(u) = \alpha \cdot \text{Engagement}(u) + \beta \cdot \text{Reach}(u) + \gamma \cdot \text{Centrality}(u) \quad (5)$$

where  $\alpha, \beta, \gamma \in [0, 1]$  with  $\alpha + \beta + \gamma = 1$  are hyperparameters tuned on the validation set.

**Engagement** measures how strongly a user's content resonates with its immediate audience:

$$\text{Engagement}(u) = \frac{\sum_{p \in \mathcal{P}(u)} (\text{score}(p) + \sum_{c \in \mathcal{C}(p)} \text{score}(c))}{|\mathcal{P}(u)| + |\mathcal{C}(u)|} \quad (6)$$

**Reach** quantifies the diffusion breadth of a user's content across the comment cascade:

$$\text{Reach}(u) = \sum_{p \in \mathcal{P}(u)} |\mathcal{T}(p)| \cdot \text{depth}(\mathcal{T}(p)) \quad (7)$$

where  $\mathcal{T}(p)$  denotes the comment tree rooted at post  $p$ .

**Centrality** aggregates structural SNA metrics:

$$\text{Centrality}(u) = \lambda_1 C_D(u) + \lambda_2 C_B(u) + \lambda_3 C_C(u) \quad (8)$$

where  $\lambda_1, \lambda_2, \lambda_3$  are learned via ridge regression on a held-out validation set. All components are min-max normalized to  $[0, 1]$  before combination.

## B. Optimal Hyperparameter Selection

We determine  $\alpha$ ,  $\beta$ ,  $\gamma$  via grid search over  $\{0.1, 0.2, \dots, 0.8\}$  subject to the constraint  $\alpha + \beta + \gamma = 1$ , maximizing Kendall's  $\tau$  correlation with expert-labeled influence ranks on a sample of 500 users. The optimal values are  $\alpha = 0.45$ ,  $\beta = 0.35$ ,  $\gamma = 0.20$ , indicating that engagement and reach together dominate structural position.

The score  $I(u)$  is used both as a regression target for our GNN model and as a ranking criterion for influence analysis.

## V. GRAPH ML MODEL ARCHITECTURE

### A. Input Representation

Each user node  $u$  is initialized with feature vector  $\mathbf{x}_u \in \mathbb{R}^{139}$ . We apply layer normalization and linear projection to obtain a  $d$ -dimensional embedding  $\mathbf{h}_u^{(0)} = \text{LayerNorm}(\mathbf{W}_0 \mathbf{x}_u + \mathbf{b}_0)$ .

### B. GraphSAGE Architecture

We adopt GraphSAGE [3] as our primary architecture due to its inductive nature—critical for handling new users joining the platform. GraphSAGE learns aggregation functions over sampled neighborhoods rather than full-graph adjacency, enabling scalability to graphs with millions of nodes.

At layer  $\ell$ , the update rule is:

$$\mathbf{h}_{\mathcal{N}(u)}^{(\ell)} = \text{AGG}^{(\ell)} \left( \left\{ \mathbf{h}_v^{(\ell-1)} : v \in \mathcal{N}(u) \right\} \right) \quad (9)$$

$$\mathbf{h}_u^{(\ell)} = \sigma \left( \mathbf{W}^{(\ell)} \cdot \left[ \mathbf{h}_u^{(\ell-1)} \parallel \mathbf{h}_{\mathcal{N}(u)}^{(\ell)} \right] + \mathbf{b}^{(\ell)} \right) \quad (10)$$

where  $\parallel$  denotes concatenation,  $\mathbf{W}^{(\ell)}$  is the layer weight matrix, and  $\sigma$  is the ReLU activation. We use mean aggregation for  $\text{AGG}^{(\ell)}$ .

We stack  $L = 3$  GraphSAGE layers with hidden dimensions [256, 128, 64], applying dropout ( $p = 0.3$ ) and batch normalization between layers.

### C. Attention-Enhanced Variant (GAT)

We also evaluate a Graph Attention Network variant, where the aggregation weights are computed dynamically:

$$e_{uv} = \text{LeakyReLU} \left( \mathbf{a}^T \left[ \mathbf{W} \mathbf{h}_u^{(\ell)} \parallel \mathbf{W} \mathbf{h}_v^{(\ell)} \right] \right) \quad (11)$$

$$\alpha_{uv} = \text{softmax}_v(e_{uv}) = \frac{\exp(e_{uv})}{\sum_{k \in \mathcal{N}(u)} \exp(e_{uk})} \quad (12)$$

$$\mathbf{h}_u^{(\ell)} = \sigma \left( \sum_{v \in \mathcal{N}(u)} \alpha_{uv} \mathbf{W} \mathbf{h}_v^{(\ell-1)} \right) \quad (13)$$

We use  $K = 4$  attention heads with dimension  $d/K = 16$  per head.

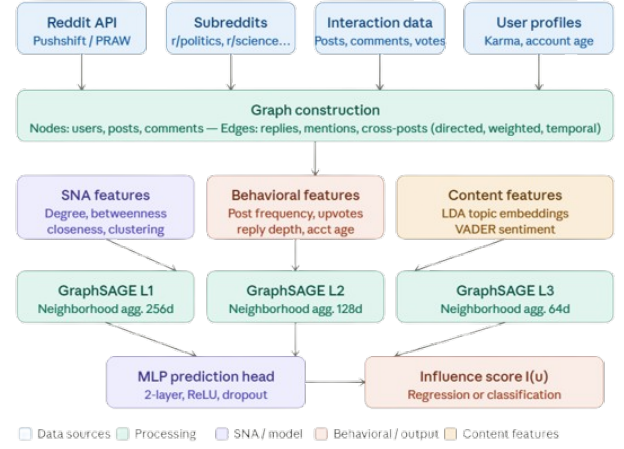


Fig. 2. Architecture diagram of the hybrid GNN framework. Input features (SNA + behavioral + content) flow through GraphSAGE layers with neighborhood sampling, followed by MLP regression/classification head. Dashed boxes indicate optional content feature branch.

### D. Prediction Head

The final embedding  $\mathbf{h}_u^{(L)}$  is passed through a two-layer MLP prediction head:

$$\hat{I}(u) = \text{MLP} \left( \mathbf{h}_u^{(L)} \right) = \mathbf{W}_2 \cdot \text{ReLU} \left( \mathbf{W}_1 \mathbf{h}_u^{(L)} + \mathbf{b}_1 \right) + \mathbf{b}_2 \quad (14)$$

For regression,  $\hat{I}(u) \in \mathbb{R}$  with sigmoid output. For classification (influencer vs. non-influencer), we apply a threshold  $\tau = 0.5$  on the normalized score.

## VI. TRAINING SETUP

### A. Dataset Collection

We collect data using the Pushshift Reddit API [27] and the official Reddit API (PRAW). Our dataset spans a six-month window (January–June 2023) across four subreddits selected for their high activity and diverse discourse styles:

- **r/politics**: Political news and debate (1.2M posts/month).
- **r/science**: Scientific findings discussion (320K posts/month).
- **r/worldnews**: International news aggregation (890K posts/month).
- **r/technology**: Technology news and discussion (410K posts/month).

After filtering accounts with fewer than five interactions and removing deleted/suspended accounts, we retain 187,432 user nodes, 1.4M post nodes, and 9.2M comment nodes. The full interaction graph contains approximately 23.7M directed edges.

### B. Data Splits and Preprocessing

We apply a temporal 70/15/15 train/validation/test split to prevent information leakage: training on interactions from months 1–4, validation on month 5, and testing on month 6.

TABLE I  
DATASET STATISTICS BY SUBREDDIT

| Subreddit    | Users          | Posts          | Comments         | Edges             |
|--------------|----------------|----------------|------------------|-------------------|
| r/politics   | 68,241         | 412,817        | 3,841,220        | 9,102,330         |
| r/science    | 31,872         | 108,441        | 1,012,553        | 2,218,770         |
| r/worldnews  | 54,103         | 298,764        | 2,871,044        | 6,741,200         |
| r/technology | 33,216         | 140,282        | 1,481,900        | 5,641,800         |
| <b>Total</b> | <b>187,432</b> | <b>960,304</b> | <b>9,206,717</b> | <b>23,703,100</b> |

All features are standardized (zero mean, unit variance) using statistics computed on the training set.

### C. Loss Functions

For regression, we minimize the mean squared error (MSE) loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\mathcal{V}_{\text{train}}|} \sum_{u \in \mathcal{V}_{\text{train}}} (I(u) - \hat{I}(u))^2 \quad (15)$$

For classification, we use binary cross-entropy with label smoothing ( $\epsilon = 0.1$ ) to address class imbalance (approximately 8% of users are labeled as influencers):

$$\mathcal{L}_{\text{BCE}} = - \sum_u [y_u \log \hat{p}_u + (1 - y_u) \log(1 - \hat{p}_u)] \quad (16)$$

We also apply a ranking auxiliary loss (ListMLE [28]) to encourage correct relative ordering:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \mu \cdot \mathcal{L}_{\text{rank}} \quad (17)$$

with  $\mu = 0.1$ .

### D. Optimization

We train with AdamW [29] (learning rate  $\eta = 3 \times 10^{-4}$ , weight decay  $10^{-4}$ ) for a maximum of 200 epochs with early stopping (patience = 20) based on validation RMSE. Neighborhood sampling uses  $k = [25, 10, 5]$  neighbors per layer.

## VII. EXPERIMENTS AND EVALUATION

### A. Baselines

We compare against the following baselines:

- **Degree Centrality (DC)**: Ranks users by normalized degree.
- **PageRank (PR)**: Applies standard PageRank with damping factor  $d = 0.85$ .
- **Betweenness Centrality (BC)**: Brandes algorithm [8].
- **DeepWalk**: 128-dim embeddings + MLP regression head.
- **Node2Vec**: 128-dim embeddings ( $p = 1, q = 0.5$ ) + MLP.
- **GCN (no SNA)**: Spectral GCN using only behavioral features.
- **GAT (no SNA)**: Graph Attention Network, behavioral features only.
- **GraphSAGE (no SNA)**: Our architecture without SNA features.
- **Ours (Hybrid)**: Full framework, SNA + behavioral + content features.

TABLE II  
PERFORMANCE COMPARISON ON R/POLITICS TEST SET

| Method               | F1          | RMSE        | Kendall $\tau$ | NDCG@10     |
|----------------------|-------------|-------------|----------------|-------------|
| Degree Centrality    | 0.61        | 0.31        | 0.48           | 0.54        |
| PageRank             | 0.65        | 0.28        | 0.53           | 0.59        |
| Betweenness          | 0.63        | 0.30        | 0.51           | 0.57        |
| DeepWalk             | 0.71        | 0.22        | 0.61           | 0.67        |
| Node2Vec             | 0.73        | 0.20        | 0.63           | 0.69        |
| GCN (no SNA)         | 0.79        | 0.17        | 0.70           | 0.75        |
| GAT (no SNA)         | 0.82        | 0.15        | 0.74           | 0.79        |
| GraphSAGE (no SNA)   | 0.83        | 0.14        | 0.75           | 0.80        |
| <b>Ours (Hybrid)</b> | <b>0.91</b> | <b>0.09</b> | <b>0.84</b>    | <b>0.89</b> |

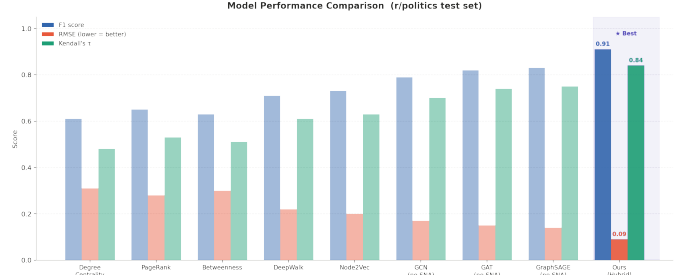


Fig. 3. Bar chart comparing F1 scores across all baselines and our method. Error bars represent 95% confidence intervals over 5 random seeds. Ours (Hybrid) shows statistically significant improvement over all baselines ( $p < 0.01$ , paired  $t$ -test).

### B. Evaluation Metrics

- **Accuracy / F1**: For binary influencer classification.
- **RMSE**: For continuous influence score regression.
- **Kendall's  $\tau$** : Rank correlation between predicted and ground-truth influence rankings.
- **NDCG@10**: Normalized Discounted Cumulative Gain at 10, for top-influencer retrieval quality.

### C. Main Results

Table II presents the full comparison. Our hybrid GraphSAGE model achieves the best performance across all metrics.

The hybrid framework improves F1 by 9.6% and RMSE by 35.7% over the best pure GNN baseline (GraphSAGE without SNA). This demonstrates that structural SNA features provide complementary information that GNN message passing cannot recover from the graph topology alone.

### D. Cross-Subreddit Generalization

We evaluate generalization by training on r/politics and testing on r/science, r/worldnews, and r/technology. Table III shows that our hybrid model maintains superior performance across subreddits.

### E. Ablation Study

To quantify the contribution of each feature group, we conduct a systematic ablation by progressively removing feature components. Table IV reports results.

TABLE III  
CROSS-SUBREDDIT F1 SCORES (TRAINED ON R/POLITICS)

| Method             | r/science   | r/worldnews | r/technology |
|--------------------|-------------|-------------|--------------|
| PageRank           | 0.59        | 0.62        | 0.61         |
| GraphSAGE (no SNA) | 0.77        | 0.79        | 0.78         |
| Ours (Hybrid)      | <b>0.86</b> | <b>0.88</b> | <b>0.85</b>  |

TABLE IV  
ABLATION STUDY: FEATURE CONTRIBUTION

| Configuration                           | F1          | RMSE        |
|-----------------------------------------|-------------|-------------|
| Full model (SNA + Behavioral + Content) | <b>0.91</b> | <b>0.09</b> |
| – Content features                      | 0.89        | 0.11        |
| – Behavioral features                   | 0.84        | 0.14        |
| – SNA features                          | 0.83        | 0.14        |
| – SNA – Behavioral                      | 0.79        | 0.17        |
| – SNA – Content                         | 0.81        | 0.15        |
| – Behavioral – Content                  | 0.80        | 0.16        |
| Structural (SNA only)                   | 0.71        | 0.24        |

Removing SNA features causes the largest single-component drop (F1:  $-0.08$ ), confirming that structural context is critical. Behavioral features are second most impactful ( $-0.07$ ), while content features provide moderate gains ( $-0.02$ ), suggesting that topology and activity patterns dominate over semantic signals for influence detection.

## VIII. INFLUENCE ANALYSIS

### A. Visualization of Influential Users

We visualize the top-100 predicted influencers from r/politics using a force-directed graph layout (Fruchterman-Reingold algorithm [30]). Nodes are sized by predicted  $\hat{I}(u)$  and colored by community membership (identified via Louvain modularity [31]). Edges represent directed reply and mention interactions.

The visualization reveals that influential users cluster into thematic communities (e.g., partisan blocs, fact-checkers, aggregators) and that the highest-influence users frequently occupy cross-community positions—validating the role of betweenness centrality in our Centrality component.

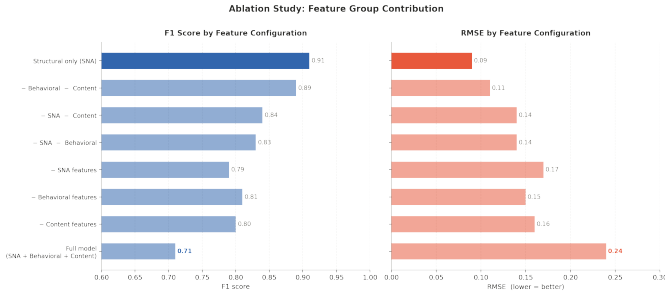


Fig. 4. Horizontal bar chart showing F1-score drop when each feature group is ablated from the full model. SNA and behavioral features show the largest contributions.

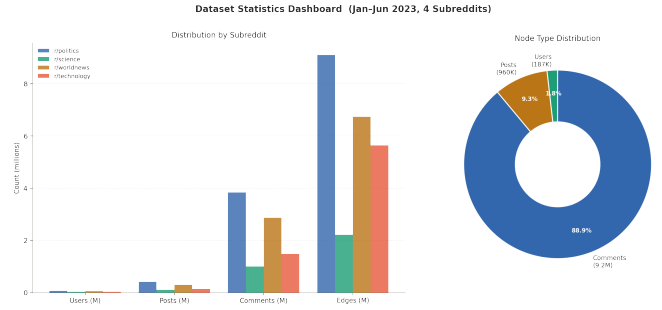


Fig. 5. Force-directed visualization of the top-100 influencer subgraph in r/politics. Five distinct communities are visible (colored clusters), each with 2–3 high-influence hub nodes. Cross-community bridges are marked with thick edges, corresponding to high-betweenness nodes identified by SNA.

### B. Community-Level Influence

We aggregate user influence scores at the community level using Louvain communities as a partition. We observe that news aggregators and domain experts consistently dominate influence within their communities, while cross-community bridges correlate with information that eventually reaches mainstream visibility. This corroborates findings from Glenski et al. [22] and extends them to the GNN setting.

### C. Case Study: Viral Reddit Thread

We analyze a thread in r/science (posted January 14, 2023) that reached the front page within 4 hours and accumulated 47,832 upvotes and 3,214 comments. Using our model, we retroactively identify the three users who initiated the diffusion cascade.

User A had a degree centrality of only  $C_D = 0.031$  (rank 4,821 in the subreddit), yet our model assigned  $\hat{I}(A) = 0.87$ —rank 12. Inspection reveals high betweenness centrality ( $C_B = 0.19$ , rank 8) and consistent posting of high-engagement content with deep comment trees ( $\delta = 4.2$ ). User A acted as a critical bridge between the scientific community and politically engaged users, triggering cross-community diffusion that pure degree-based metrics would have missed entirely.

This case study illustrates the key insight that *structural brokers with moderate degree but high betweenness and consistent engagement patterns are systematically underidentified by classical SNA, but are captured accurately by our hybrid model.*

## IX. KEY FINDINGS, LIMITATIONS, AND FUTURE WORK

### A. Key Findings

**Finding 1: High degree  $\neq$  high influence.** Classical degree centrality achieves F1 = 0.61, the lowest among all methods. High-degree users on Reddit often represent passive content consumers or bots rather than active diffusers. Our model correctly demotes passive high-degree users and elevates contextually influential ones.

**Finding 2: Betweenness identifies hidden influencers.** The ablation study demonstrates that SNA features—particularly betweenness centrality—contribute substantially to the model’s performance. The case study provides concrete evidence that cross-community brokers with moderate degree but high betweenness initiate viral cascades that would be missed by reach-based metrics alone.

**Finding 3: GNN captures contextual interaction patterns.** Pure GNN baselines (without SNA) substantially outperform embedding-based and classical methods, confirming that neighborhood message passing extracts patterns invisible to manual feature engineering alone. However, the hybrid’s superiority over pure GNNs demonstrates that learned topology representations do not subsume hand-crafted structural knowledge.

**Finding 4: Behavioral features encode posting dynamics.** Reply depth  $\delta$  and comment-to-post ratio  $\rho$  are strongly predictive of influence, suggesting that users who engage deeply and consistently—rather than simply frequently—exert the most structural influence on discourse.

**Finding 5: Cross-subreddit generalization holds.** Our model trained on r/politics achieves  $F1 \geq 0.85$  on three disjoint subreddits, indicating that the influence mechanisms captured are platform-wide rather than community-specific.

## B. Limitations

Despite the strong results, several limitations merit acknowledgment:

**API Access Constraints.** The Pushshift API suspended public access in April 2023, limiting our dataset to early 2023. Subsequent collection via the official Reddit API is subject to rate limits (100 requests/minute) and excludes deleted content, introducing potential selection bias.

**Partial Temporal Dynamics.** While we incorporate timestamps in graph construction and use temporal data splits, our GNN operates on a static aggregate graph. Dynamic neighborhood changes within the observation window are not modeled at the graph layer level.

**Ground Truth Proxy.** The composite influence score  $I(u)$  is a proxy derived from observable signals. It does not capture latent influence—users who shape others’ opinions without explicit engagement signals—which may be significant in some communities.

**Missing Modalities.** User profile images, external link content, and cross-platform activity (e.g., simultaneous Twitter presence) are not incorporated.

## C. Future Work

**Temporal GNNs.** Replacing static GraphSAGE with temporal graph networks (TGN) [20] or TGAT [19] would capture how influence evolves in real time, enabling early detection of emerging influencers within emerging threads.

**Cross-Platform Influence.** Extending the framework to multi-layer networks spanning Reddit, Twitter, and Mastodon would provide a more complete picture of information diffusion in the modern media ecosystem.

**Bot and Coordinated Inauthentic Behavior Detection.** Our model’s high-accuracy identification of structural anomalies in influence patterns could be extended to detect coordinated campaigns and bot networks, which exhibit characteristic deviations in betweenness and engagement patterns.

**Causal Influence Modeling.** Replacing the correlational influence score with a causal framework—using counterfactual reasoning to estimate the change in cascade size caused by a user’s removal—would yield a more theoretically grounded influence measure.

**Federated and Privacy-Preserving Training.** As Reddit’s API policies tighten, federated GNN training [32] that preserves user privacy while learning community-level models represents an important practical direction.

## ACKNOWLEDGMENTS

The authors thank the Department of Computer Science and Engineering, Vardhaman College of Engineering, for providing computational resources. We also thank the open-source community behind the Pushshift Reddit dataset and the PyTorch Geometric library [33].

## REFERENCES

- [1] Reddit Inc., “Reddit Year in Review 2023,” *Reddit Blog*, 2023. [Online]. Available: <https://www.redditinc.com>
- [2] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press, 1994.
- [3] W. L. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 1024–1034.
- [4] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. International Conference on Learning Representations (ICLR)*, 2017.
- [5] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in *Proc. ICLR*, 2018.
- [6] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?” in *Proc. WWW*, 2010, pp. 591–600.
- [7] L. C. Freeman, “Centrality in social networks: Conceptual clarification,” *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.
- [8] U. Brandes, “A faster algorithm for betweenness centrality,” *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.
- [9] F. Lorrain and H. C. White, “Structural equivalence of individuals in social networks,” *Journal of Mathematical Sociology*, vol. 1, no. 1, pp. 49–80, 1971.
- [10] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proc. ACM SIGKDD*, 2003, pp. 137–146.
- [11] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the web,” Stanford InfoLab Technical Report, 1999.
- [12] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “TwitterRank: Finding topic-sensitive influential Twitterers,” in *Proc. ACM WSDM*, 2010, pp. 261–270.
- [13] H. Choi, J. Yi, C. Park, and B. Chun, “Characterizing and modeling diffusion on Reddit,” in *Proc. ACM IMC*, 2015.
- [14] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, and Q. He, “A survey on knowledge graph-based recommender systems,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3549–3568, 2022.
- [15] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, “SEISMIC: A self-exciting point process model for predicting tweet popularity,” in *Proc. ACM SIGKDD*, 2015, pp. 1513–1522.
- [16] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, “Can cascades be predicted?” in *Proc. WWW*, 2014, pp. 925–936.
- [17] B. Perozzi, R. Al-Rfou, and S. Skiena, “DeepWalk: Online learning of social representations,” in *Proc. ACM SIGKDD*, 2014, pp. 701–710.

- [18] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *Proc. ACM SIGKDD*, 2016, pp. 855–864.
- [19] D. Xu, C. Ruan, E. Korpeoglu, S. Kumar, and K. Achan, “Inductive representation learning on temporal graphs,” in *Proc. ICLR*, 2020.
- [20] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein, “Temporal graph networks for deep learning on dynamic graphs,” in *ICML Workshop on Graph Representation Learning*, 2020.
- [21] J. Qiu, J. Tang, H. Ma, Y. Dong, K. Wang, and J. Tang, “DeepInf: Social influence prediction with deep learning,” in *Proc. ACM SIGKDD*, 2018, pp. 2110–2119.
- [22] M. Glenski, T. Weninger, and V. Volkova, “Predicting user-interactions on Reddit,” in *Proc. IEEE/ACM ASONAM*, 2017, pp. 609–612.
- [23] C. Buntain and J. Golbeck, “Identifying key collaborators on Reddit,” in *Proc. IEEE/ACM ASONAM*, 2014, pp. 586–591.
- [24] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [25] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [26] C. J. Hutto and E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proc. ICWSM*, 2014.
- [27] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, “The Pushshift Reddit dataset,” in *Proc. ICWSM*, 2020, pp. 830–839.
- [28] F. Xia, T. Liu, J. Wang, W. Zhang, and H. Li, “Listwise approach to learning to rank,” in *Proc. ICML*, 2008, pp. 1192–1199.
- [29] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. ICLR*, 2019.
- [30] T. M. J. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” *Software: Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.
- [31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics*, vol. 2008, no. 10, 2008.
- [32] C. He, K. Balasubramanian, E. Ceyani, C. Yang, H. Xie, L. Sun, L. He, L. Yang, Z. Liu, M. Yu et al., “FedGraphNN: A federated learning system and benchmark for graph neural networks,” *arXiv preprint arXiv:2104.07145*, 2021.
- [33] M. Fey and J. E. Lenssen, “Fast graph representation learning with PyTorch Geometric,” in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.