Webpage

Code & Data

Predicting Information Pathways Across Online Communities

Yiqiao Jin[1], Yeon-Chang Lee[1], Kartik Sharma[1], Meng Ye[2],
Karan Sikka[2], Ajay Divakaran[2], Srijan Kumar[1]
[1]Georgia Institute of Technology
[2]SRI International

https://github.com/claws-lab/INPAC

KDD2023
AUGUST 6-10
29TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING

SRI International

Georgia Tech

Outline

- Introduction
- Preliminary
- Method
- Evaluation

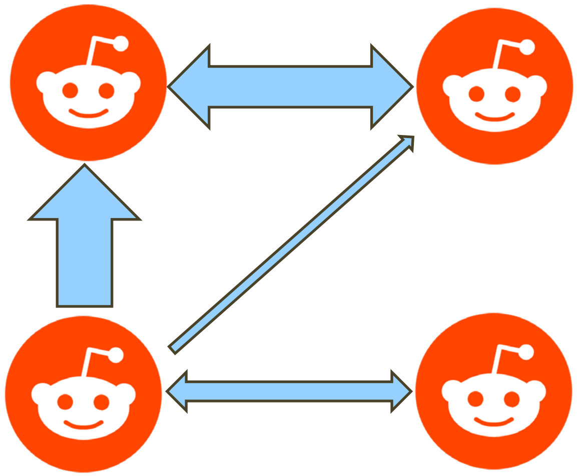
Webpage

Code & Data

Introduction

Information Pathways

- Social media users form communities based on their interests, beliefs, ethnicity, and geographical location
- These communities interact with & influence one another
- The underlying pathways on which information propagates remain relatively stable.



CLIPP

- The problem of Community-level Information Pathway Prediction (**CLIPP**) aims at predicting the transmission trajectory of content across online communities.
- Importance
  - Facilitates the distribution of **valuable information** to a larger audience
  - Prevents the proliferation of **harmful information.**

Travel guide

Challenges

- Inter-community relationships and influence are unknown
- Information spread is multi-modal
- New content and new communities appear over time.

This Work

- We investigate the dynamics of community-level information flow while jointly addressing the challenges of complex diffusion environment and the continuously evolving information ecosystem.
- We investigate YouTube videos shared on Reddit
  - YouTube videos contain multimodal information, including text (title and descriptions), visual (videos and thumbnail images), and channel information
  - Reddit is characterized by its numerous communities named "**subreddits**"
  - Each **subreddit** is dedicated to specific topics or interests
  - Ideal for studying <u>community-level</u> information spread

Preliminary

Preliminary

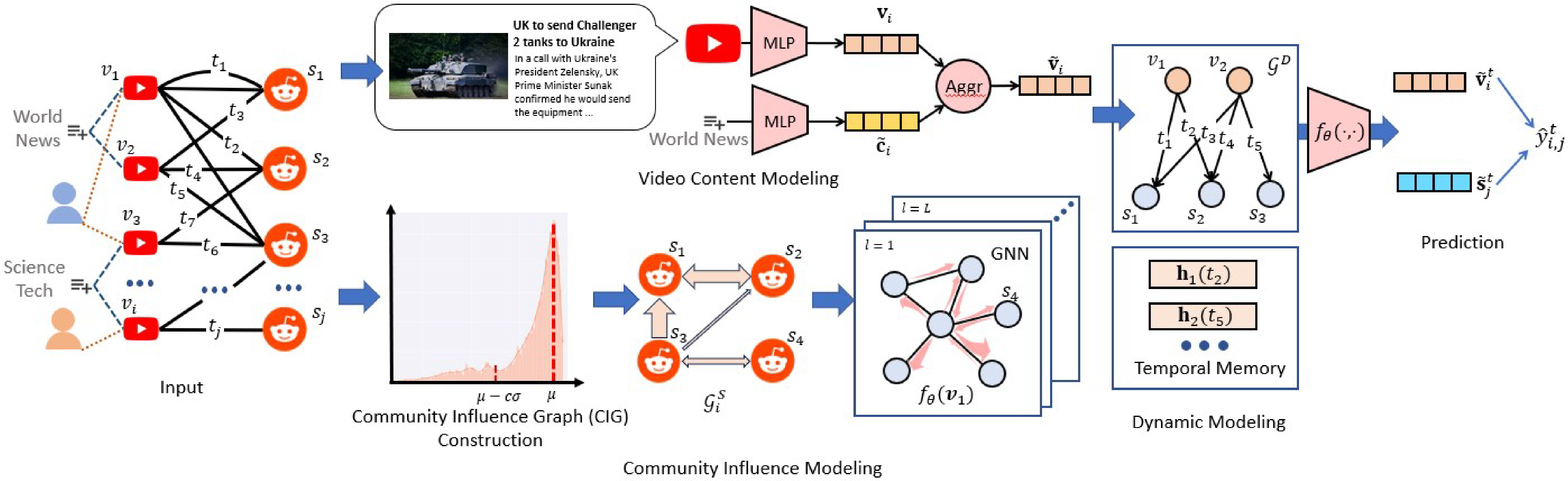
- A posting of a video means a video link appearing on a subreddit, either as a standalone post or as part of a longer post.
- A posting instance is a 4-tuple $p_{ij} = (v_i, s_j, u_j, t_j)$, where v_i is a video posted by a user u_j in an online community s_j at time t_j.

PROBLEM 1 (INFORMATION PATHWAY PREDICTION). *Given a video v_i, its posting sequence $P_i = \{(v_i, s_j, u_j, t_j)\}_{j=1}^{N}$ with length N, and a target timestamp $t_{j'}$, our model outputs a ranked list of communities $\{s_k\}$ indicating the most likely communities that v_i will appear at time $t_{j'}$.*

Georgia Tech

Method

The Proposed Framework - INPAC



Our Framework *Information Pathway Across Online Communities* **(INPAC)** consists of static modeling, including video content and community influence modeling, as well as dynamic modeling.

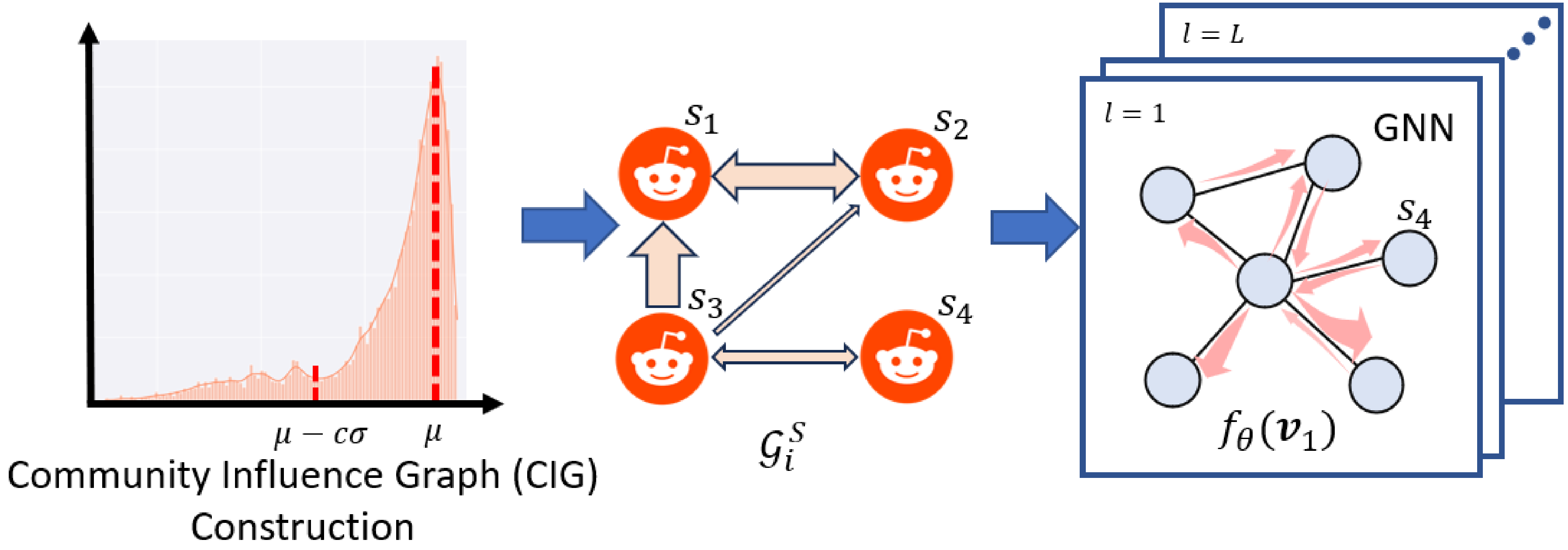
Georgia Tech

Static Modeling

Title of the Video	Subreddits on Which the Video Appears
Canadian Trudeau Investigation	Liberate_Canada → conspiracy → TheNewRight → PeoplesPartyofCanada → Canada_First
Reviews: Super Dragon Ball Heroes Episode 19	promote → AnimeReviews → anime_manga → YouTubeAnimeCommunity → Anime_and_Manga
Warcraft 3 Reforged Cutscene Only	WC3 → pcgaming → warcraft3 → gaming → legaladviceofftopic
Practical Greeting Phrases for Chinese New Year	learnchinese → learnmandarin → learnmandarinchinese
Accepting what is. (Realize Instant Freedom)	AnxietyDepression → Soulnexus → SpiritualAwakening → Meditation → spirituality → awakened → inspiration
Covid-19 Explained with Data Science	Python → CoronavirusUS → CanadaCoronavirus → CoronaVirus_2019_nCoV → CoronavirusUK
Implement RNN-LSTM for Music Genre Classification	learnmachinelearning → Python → tensorflow → musictheory

- **Challenge** inter-community relationships and influence are unknown
- **Insight** a video is usually shared (by like-minded users) in topically similar communities, which can be used to infer such relations.

Community Influence Modeling

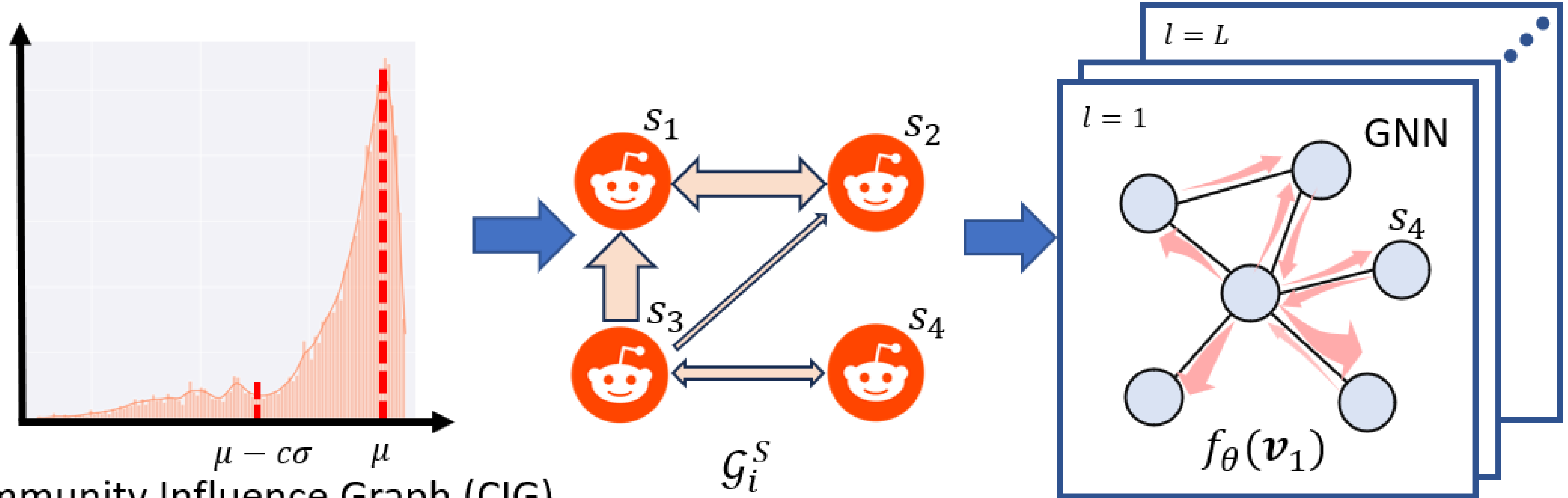


Community Influence Modeling

- **Community-level influence** exists when two communities share a common group of users.
- Such influence can be inferred through two aspects:
- **Sequential Signals** The sequence of communities $\{s_1, s_2, \ldots\}$ in which a video is posted.
- **Temporal Signals** Users require a certain amount of time to engage in online content. The interval between the appearance of v_i in s_1 and s_2 serves as an indicator of the influence of s_1 on the appearance of v_i in s_2.

Georgia Tech

Static Modeling



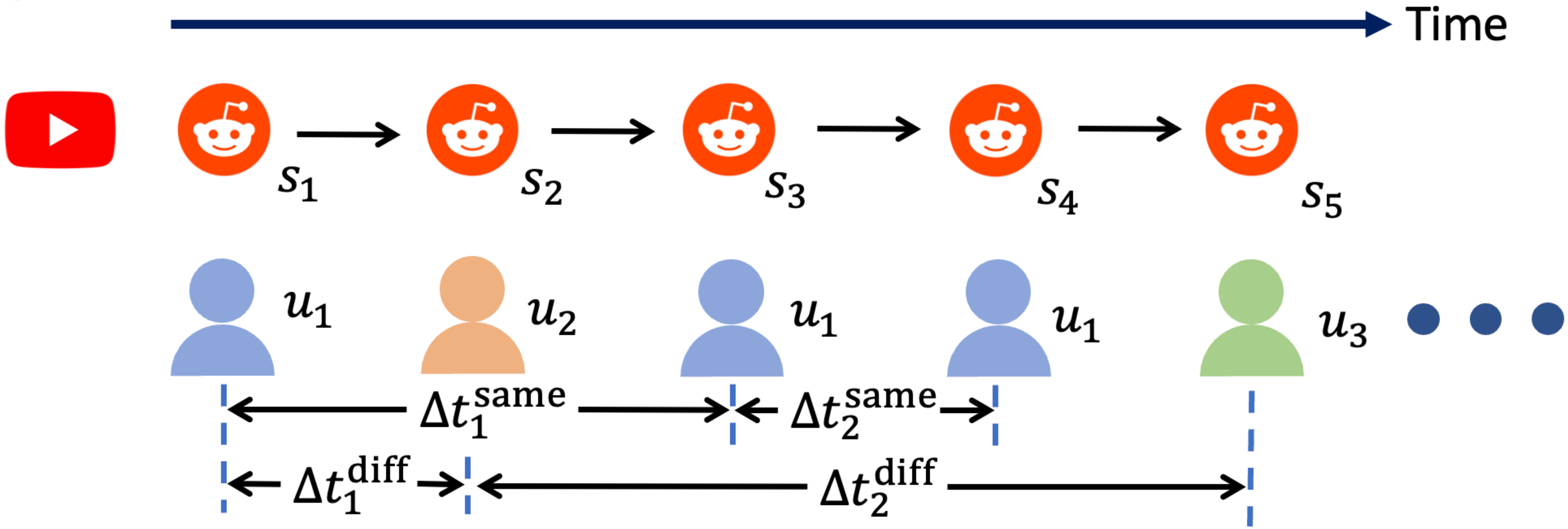
Community Influence Graph (CIG) Construction

Community Influence Modeling

We use the propagation sequence of a video to infer the community-to-community influence

Georgia Tech

Community Influence Modeling



- We calculate two types of time intervals in users' sharing behaviors
- Δt^{Same}: time intervals between consecutive shares of v_i by the same user
- Δt^{Diff} : time intervals between the first share of v_i by different users

Community Influence Modeling

- Δt^{Diff} has a unimodal distribution with mean 6.844 and stdev 0.823 on the log scale
- We determine the cutoff time for partitioning sessions using a threshold time $\Delta t^{\mathrm{Thres}}$

$$\Delta t^{\mathrm{Thres}} = \mu - c\sigma$$

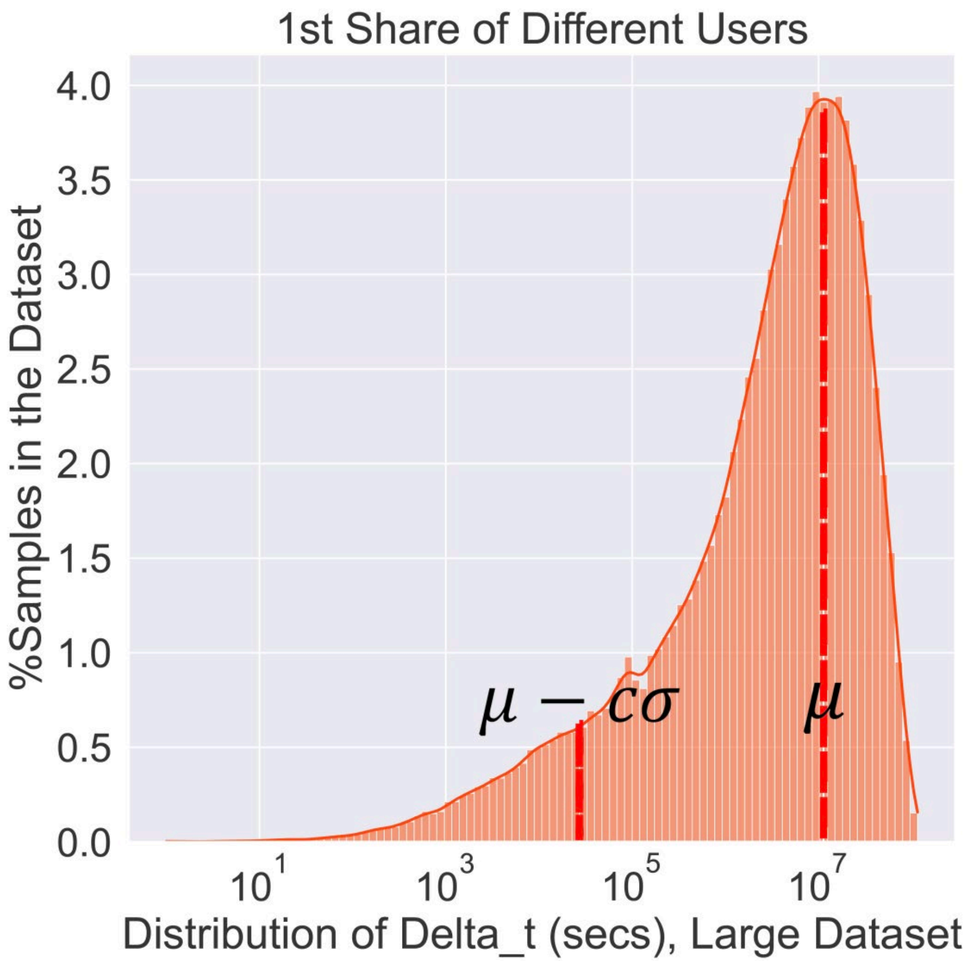
- Then, we construct the edges for the CIGs

Case of Different User

- Directed edge $s_j \rightarrow s_k$ if two shares of some video v_i from different users occur $\leq \Delta t^{\mathrm{Thres}}$
- No edge if $> \Delta t^{\mathrm{Thres}}$

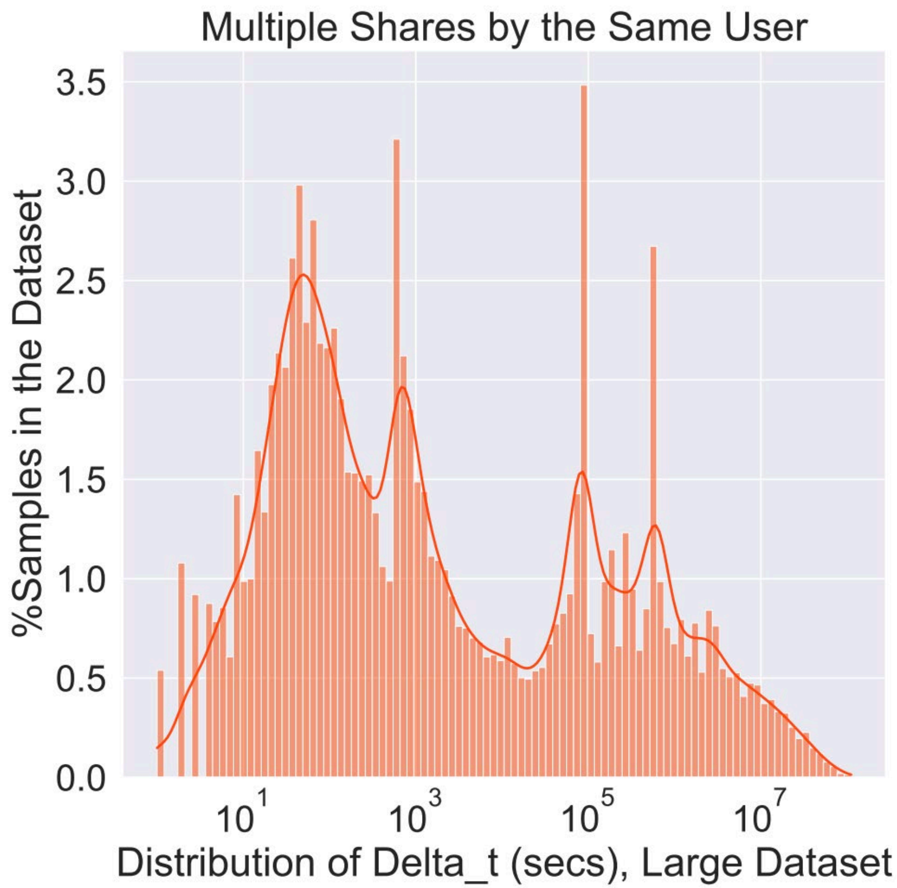
Case of Same User

- Bidirectional edge $s_j \leftrightarrow s_k$ if there exist two shares of some video v_i from the same user.
    - This is due to mutual influence in terms of content sharing.

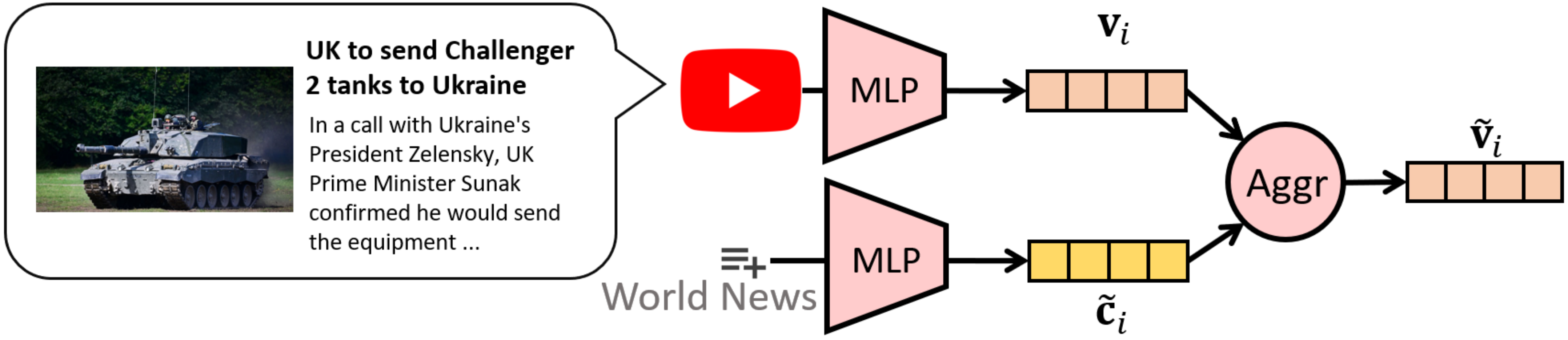


Community Influence Modeling

- We also observe that Δt^{Same} form a bimodal distribution
- A user can post the same content in various venues to enhance its visibility and attract more "likes"
- This should not be viewed as one community influencing another
- Not indicative of natural flow of content from one community to another



Video Content Modeling



Video Content Modeling

$$\tilde{\mathbf{v}}_i = \mathrm{Aggr}(\mathbf{v}_i, \mathbf{c}_{\rho(i)})$$

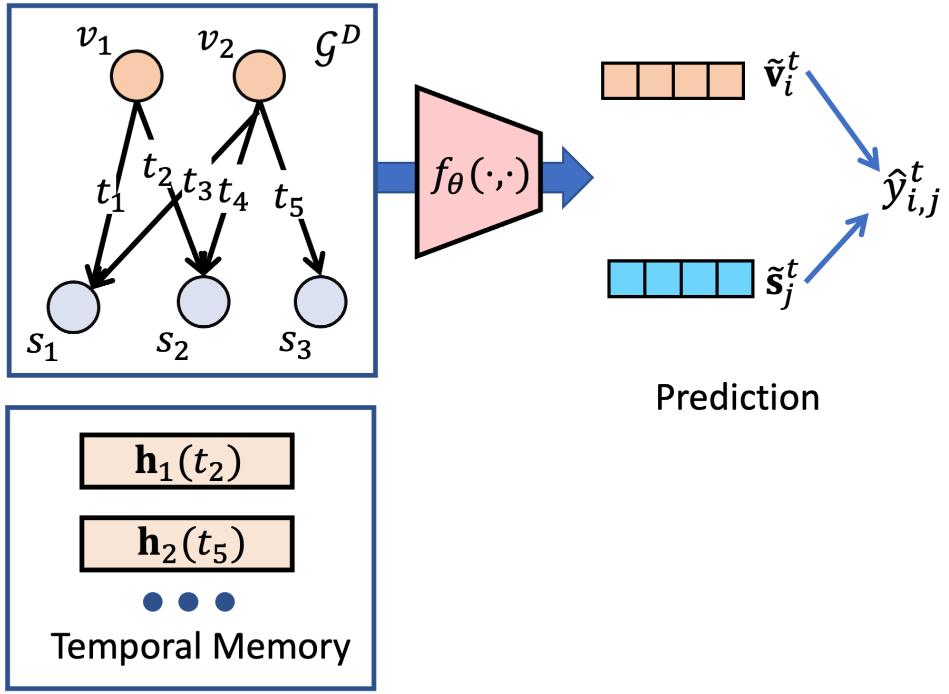
- $\mathbf{v}_i$: feature vector for the title and description of a video
- $\mathbf{c}_{\rho(i)}$: feature vector for the video's channel

Dynamic Modeling

- We leverage temporal graph network (TGN) to derive the representations $\tilde{\mathbf{v}}_i^t$ and $\tilde{\mathbf{s}}_j^t$ at time t

$$\tilde{\mathbf{v}}_i^t = f_\theta(\mathbf{h}_i(t), \mathcal{G}^D)$$
$$\tilde{\mathbf{s}}_j^t = f_\theta(\mathbf{h}_j(t), \mathcal{G}^D)$$

- $\mathbf{h}_i(t)$ and $\mathbf{h}_j(t)$ are the memory vectors for v_i and s_j, respectively.



Dynamic Modeling

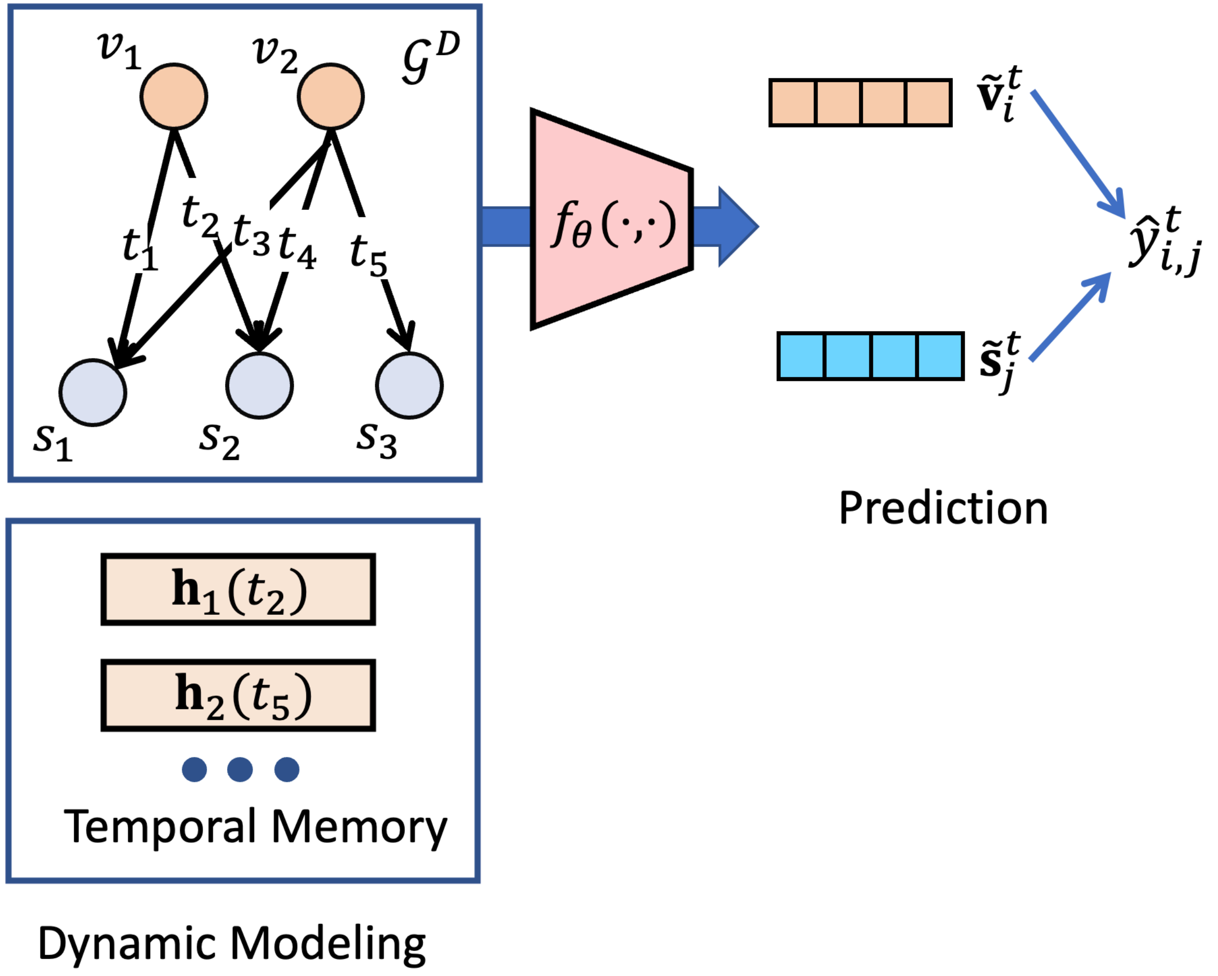
Dynamic Modeling

- We leverage temporal graph network (TGN) to derive the representations $\tilde{\mathbf{v}}_i^t$ and $\tilde{\mathbf{s}}_j^t$ at time t
- During prediction, we derive the score be- tween each video v_i and each community s_j at time t

$$\hat{y}_{ij}^t = \mathrm{MLP}(\tilde{\mathbf{v}}_i^t \odot \mathrm{MLP}(\tilde{\mathbf{s}}_j^t))$$

- The model is trained using BPR Loss

$$\mathcal{L}_{\mathrm{BPR}} = \sum_{(i,j^+,j^-,t)} -\ln(\mathrm{sigmoid}(\hat{y}_{ij^+}^t - \hat{y}_{ij^-}^t))$$



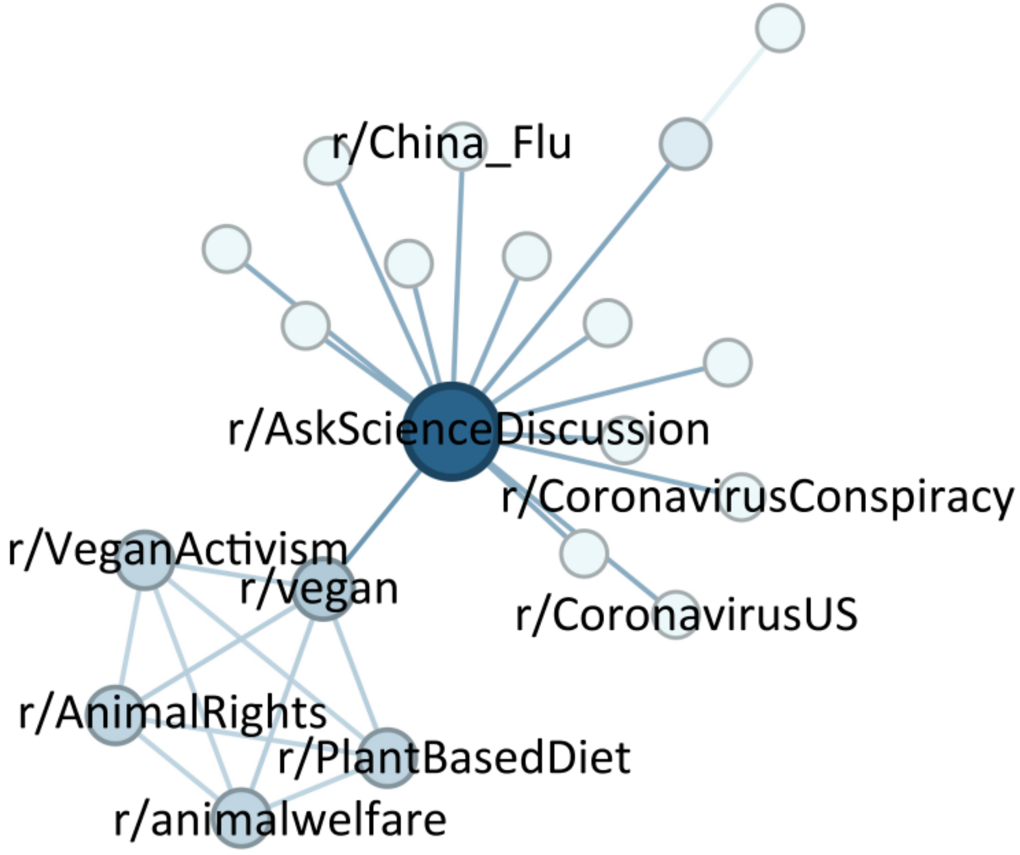
Evaluation

Results

Performance improvements of up to：

- 13.8% on NDCG@5
- 6.2% on Recall@5
- 18.8% on MRR

INPAC can identify meaningful community influence graphs (CIG) in various scenarios.



Code & Data: https://github.com/claws-lab/INPAC

Georgia Tech

Performances on Warm-start Videos

(a) Large Dataset

	Popular Communities					Non-Popular Communities				
	NDCG@5	Rec@5	NDCG@10	Rec@10	MRR	NDCG@5	Rec@5	NDCG@10	Rec@10	MRR
MF	0.5216	0.7194	0.6002	0.8469	0.4734	0.1346	0.2205	0.1820	0.3595	0.1513
NGCF	0.5291	0.7307	0.5597	0.8477	0.5246	0.1399	0.2213	0.1845	0.3680	0.1581
LightGCN	0.5468	0.7349	0.5675	0.8505	0.5215	0.1537	0.2426	0.1987	0.3832	0.1691
SVD-GCN	0.5677	0.7572	0.6002	0.8514	0.5379	0.1609	0.2539	0.2065	0.3960	0.1739
TiSASRec	0.5696	0.7593	0.6029	0.8534	0.5354	0.1668	0.2586	0.2078	0.3956	0.1770
TGAT	0.5679	0.7603	0.6130	0.8530	0.5354	0.1684	0.2590	0.2121	0.3969	0.1775
TGN	0.5723	0.7604	0.6140	0.8569	0.5576	0.1687	0.2596	0.2138	0.3970	0.1818
INPAC	**0.6013**	**0.7816**	**0.6383**	**0.8793**	**0.5822**	**0.1798**	**0.2741**	**0.2263**	**0.4182**	**0.1923**
Impr	**5.1%**	**2.8%**	**4.0%**	**3.0%**	**4.4%**	**6.6%**	**5.6%**	**5.9%**	**5.3%**	**5.8%**

(b) Small Dataset

	Popular Communities					Non-Popular Communities				
	NDCG@5	Rec@5	NDCG@10	Rec@10	MRR	NDCG@5	Rec@5	NDCG@10	Rec@10	MRR
MF	0.3594	0.5211	0.4017	0.6585	0.3356	0.0764	0.1203	0.0991	0.1958	0.0803
NGCF	0.3641	0.5282	0.4100	0.6620	0.3411	0.0807	0.1250	0.1000	0.1816	0.0887
LightGCN	0.3789	0.5493	0.4167	0.6796	0.3448	0.0852	0.1321	0.1172	0.2241	0.0967
SVD-GCN	0.3893	0.5634	0.4235	0.6839	0.3621	0.0947	0.1415	0.1204	0.2311	0.1011
TiSASRec	0.3907	0.5617	0.4287	0.6840	0.3642	0.0948	0.1439	0.1233	0.2335	0.1061
TGAT	0.3922	0.5669	0.4276	0.6845	0.3676	0.0953	0.1445	0.1256	0.2321	0.1095
TGN	0.4037	0.5728	0.4324	0.6849	0.3753	0.0981	0.1462	0.1302	0.2358	0.1156
INPAC	**0.4377**	**0.6092**	**0.4613**	**0.7031**	**0.4026**	**0.1115**	**0.1533**	**0.1428**	**0.2524**	**0.1380**
Impr.	**8.4%**	**6.3%**	**6.7%**	**2.7%**	**7.3%**	**13.6%**	**4.8%**	**9.7%**	**7.0%**	**19.4%**

Performances on Cold-start Videos

(a) Large Dataset

	Popular Communities					Non-Popular Communities				
	NDCG@5	Rec@5	NDCG@10	Rec@10	MRR	NDCG@5	Rec@5	NDCG@10	Rec@10	MRR
MF	0.5291	0.7361	0.5669	0.8411	0.4824	0.1069	0.1593	0.1390	0.2600	0.1245
NGCF	0.5632	0.7485	0.5862	0.8380	0.5118	0.1371	0.2285	0.1834	0.3732	0.1508
LightGCN	0.5768	0.7534	0.6005	0.8373	0.5247	0.1426	0.2515	0.1942	0.3926	0.1576
SVD-GCN	0.5808	0.7633	0.6033	0.8398	0.5344	0.1484	0.2532	0.1944	0.3972	0.1696
TiSASRec	0.5853	0.7604	0.6023	0.8380	0.5326	0.1516	0.2538	0.1990	0.3973	0.1705
TGAT	0.5896	0.7638	0.6104	0.8435	0.5497	0.1586	0.2549	0.2067	0.4009	0.1760
TGN	0.5872	0.7636	0.6102	0.8404	0.5452	0.1623	0.2552	0.2080	0.4019	0.1732
INPAC	**0.6174**	**0.7855**	**0.6397**	**0.8677**	**0.5776**	**0.1764**	**0.2705**	**0.2205**	**0.4238**	**0.1873**
Impr.	**4.7%**	**2.8%**	**4.8%**	**2.9%**	**5.1%**	**8.6%**	**6.0%**	**6.0%**	**5.5%**	**6.4%**

(b) Small Dataset

	Popular Communities					Non-Popular Communities				
	NDCG@5	Rec@5	NDCG@10	Rec@10	MRR	NDCG@5	Rec@5	NDCG@10	Rec@10	MRR
MF	0.3524	0.5785	0.4167	0.8508	0.2922	0.0730	0.1134	0.1077	0.2150	0.0961
NGCF	0.3631	0.5864	0.4332	0.8351	0.3194	0.0816	0.1237	0.1099	0.2320	0.0991
LightGCN	0.3958	0.5890	0.4421	0.8639	0.3221	0.0825	0.1289	0.1107	0.2262	0.0984
SVD-GCN	0.4034	0.6073	0.4515	0.8743	0.3283	0.0800	0.1289	0.1136	0.2268	0.1011
TiSASRec	0.4172	0.6466	0.4682	0.8807	0.3643	0.0849	0.1366	0.1142	0.2320	0.1071
TGAT	0.4244	0.6709	0.4779	0.8814	0.3664	0.0839	0.1392	0.1149	0.2371	0.1073
TGN	0.4273	0.6753	0.4797	0.8831	0.3696	0.0883	0.1443	0.1157	0.2396	0.1094
INPAC	**0.4646**	**0.7155**	**0.5083**	**0.9110**	**0.3847**	**0.1008**	**0.1526**	**0.1272**	**0.2506**	**0.1180**
Impr.	**8.7%**	**5.9%**	**5.9%**	**3.1%**	**4.1%**	**14.2%**	**5.8%**	**10.0%**	**4.6%**	**7.9%**

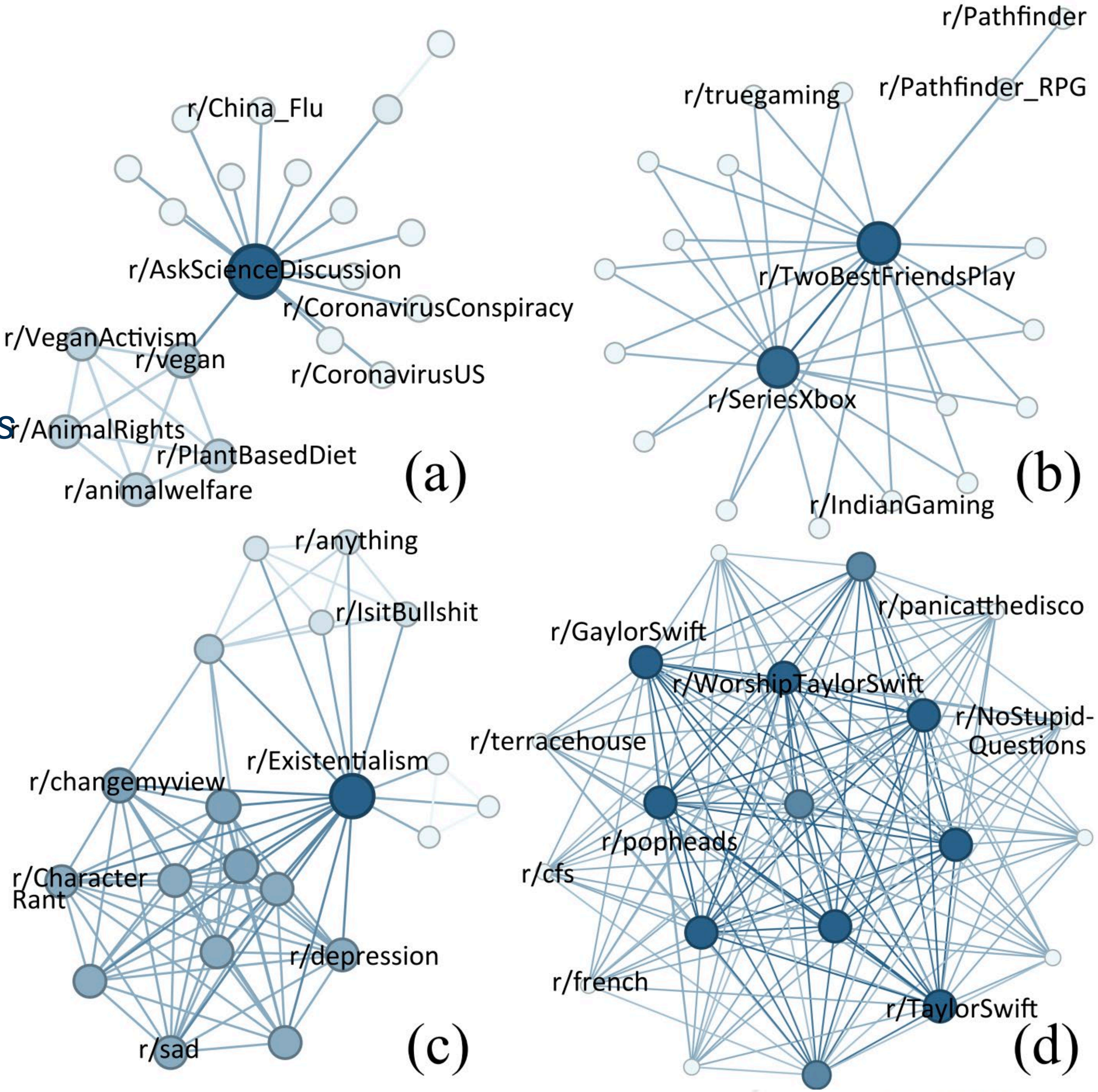
Georgia Tech

Analysis of Community Influence Graphs (CIGs)

These 4 videos were all propagated in exactly 20 communities

a. *How Wildlife Trade is Linked to Coronavirus*
b. *Black Myth: Wukong - Official 13 Minutes Gameplay Trailer*
c. *Thought experiment "BRAIN IN A VAT"*
d. *Taylor Swift - ME!*

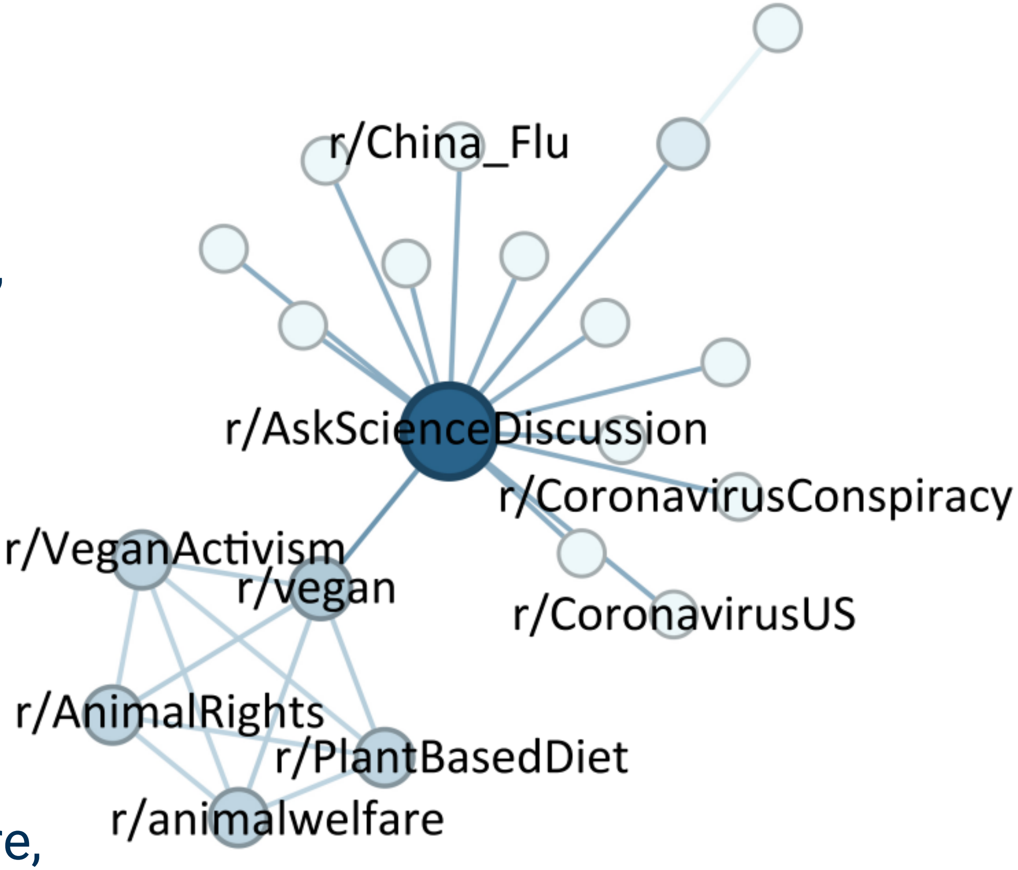
- Node sizes & colors indicate node degrees
- Edge colors indicate the edge weights
- **Observation**: CIGs generated from different videos demonstrate diverse connectivities and structures.



Analysis of Community Influence Graphs (CIGs)

(a) *How Wildlife Trade is Linked to Coronavirus* exhibits weaker connectivity with a total edge weight of 25.

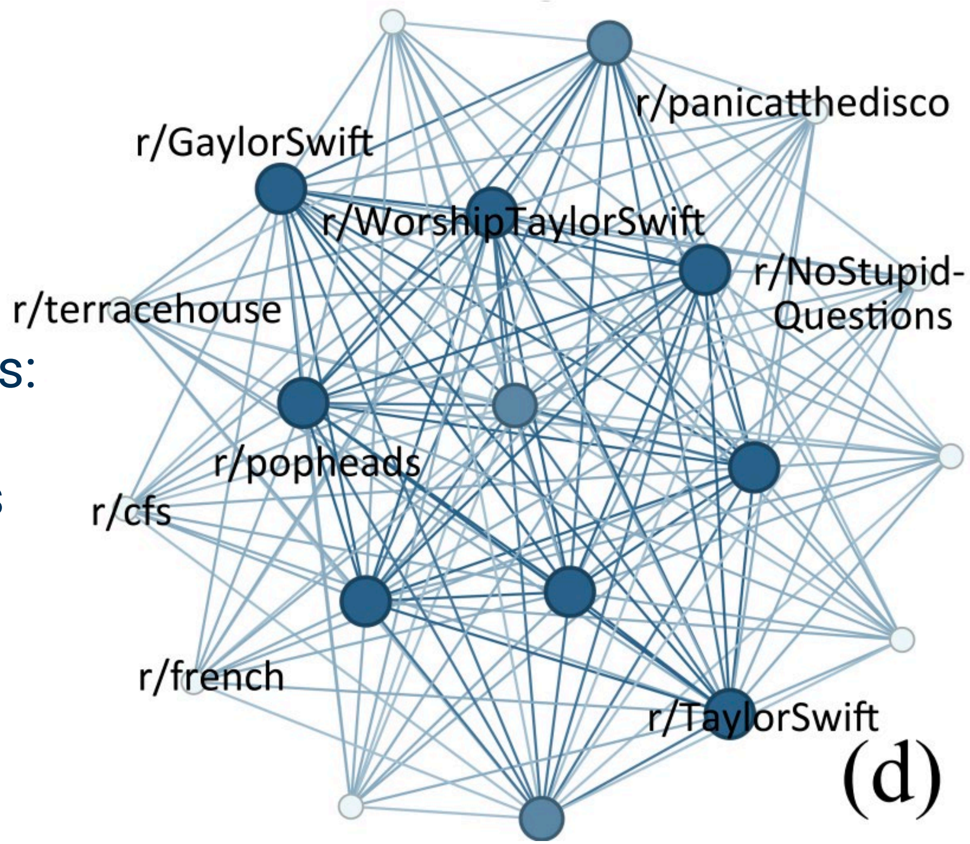
- This CIG exhibit multiple clusters
- The video was initially shared in *r/AskScienceDiscussion*, a community focused on in-depth scientific discussions, which aligned with the video's original purpose
- Then, it spread to multiple semantically similar communities within a short time OR through the same group of users
- As the video gained popularity, it was shared by distinct users in highly active COVID-19 related communities
- Eventually, the video was shared in 5 topically similar communities related to vegetarianism and animal welfare, such as *r/AnimalRights*.



Analysis of Community Influence Graphs (CIGs)

(d) *Taylor Swift - ME!*.

- The video exhibits a single cluster.
- The video first appeared in r/WorshipTaylorSwift, which directly relates to the posted video
- Subsequently, the video propagated to multiple semantically distinct communities at different time periods:
    - r/terracehouse: reality TV show
    - r/NoStupidQuestions: discussion of curious questions



Insights

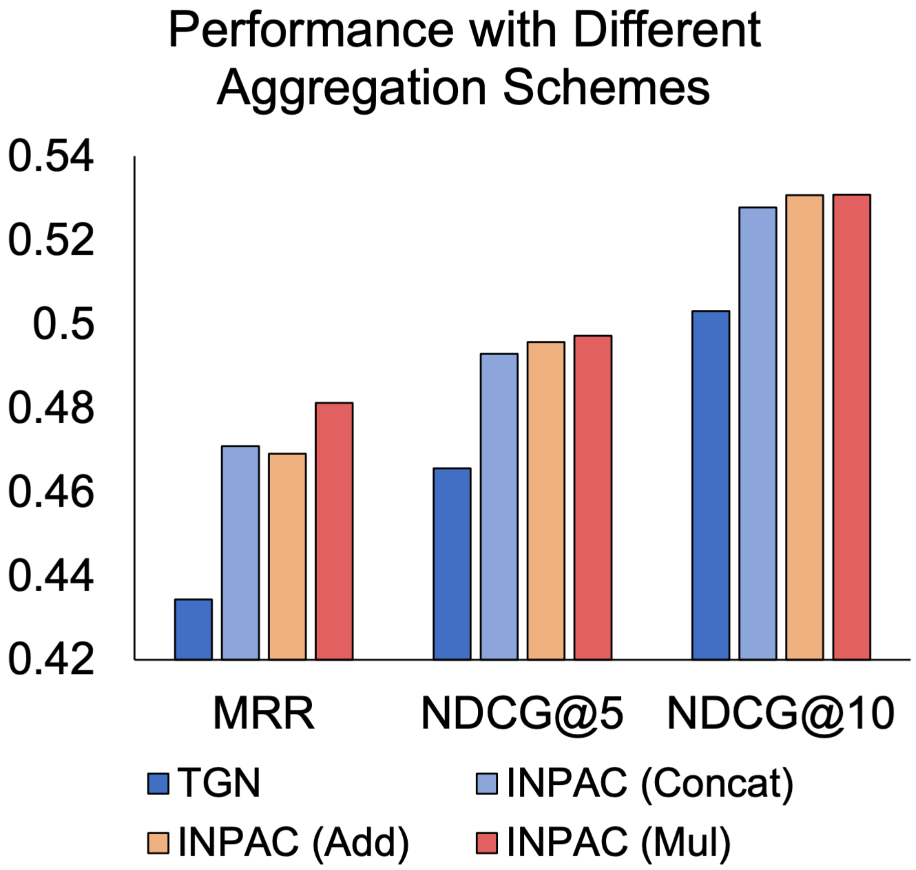
1. Initially, online content tends to be shared within communities that closely match its topic. As the content gains popularity, it gradually spreads to multiple communities with a broader range of topics.
2. Content is shared within topically similar communities in a short period, regardless of whether it is shared by the same user or different users.
3. Existence of “super spreaders” on online platforms who actively engage in and disseminate content across multiple topically diverse communities.

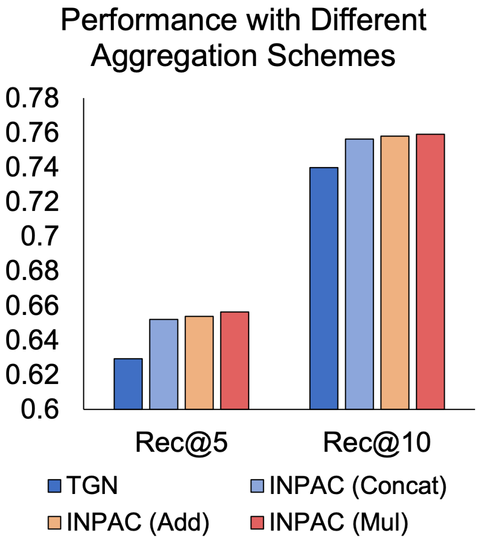
Ablation Studies

We evaluate 3 variants of INPAC
Each of them applies a separate aggregation scheme for the video & channel embeddings

- **INPAC (Add):** Addition
- **INPAC (Concat):** Concatenation
- **INPAC (Mul):** Multiplication

- **INPAC (Mul)** outperforms other variants of INPAC
- The greatest performance improvement is on MRR
- All variants outperform the strongest baseline TGN





Webpage

Code & Data

Predicting Information Pathways Across Online Communities

Yiqiao Jin[1], Yeon-Chang Lee[1], Kartik Sharma[1], Meng Ye[2], Karan Sikka[2], Ajay Divakaran[2], Srijan Kumar[1]

[1]Georgia Institute of Technology

[2]SRI International

https://github.com/claws-lab/INPAC

KDD2023

AUGUST 6-10

29TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING

SRI International

Georgia Tech