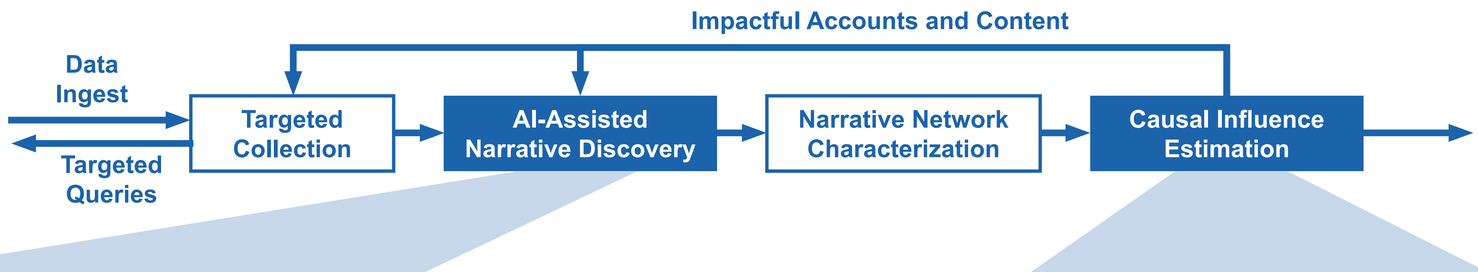
Causal Influence Pathway Quantification on Social Networks

Abstract & System Flowchart

Quantifying and characterizing the spread of narrative content and sentiment from source communities to audience communities has important applications in communication and marketing as well as in national security such as countering disinformation. This poster presents a novel method to estimate the causal influence of pathways between individuals across communities on a social network. To enable data-driven influence quantification, discovery of narrative content and aspect-based sentiment classification are automated using large language models. We demonstrate the utility of this method in identifying influential accounts both as sources and bridges of the two opposing communities competing for the attention of the audience community, on the spread of Biolabs-in-Ukraine narratives, on Twitter in 2022. Validation using tweet following statistics, external corroboration, and predictions shows that causal influence reveals hidden influencers and is more accurate than existing metrics.

Influence Quantification System

Input
Social and news media data sources

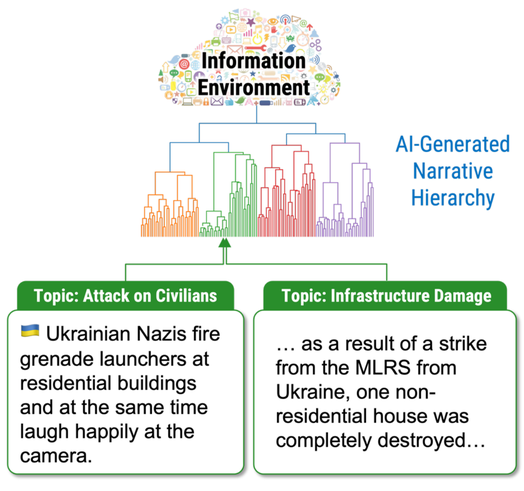


Output
- Narrative content
- Network mapping
- Causal influence
- Influence pathways
- Predictive influence

Human-AI Teaming for Narrative Discovery

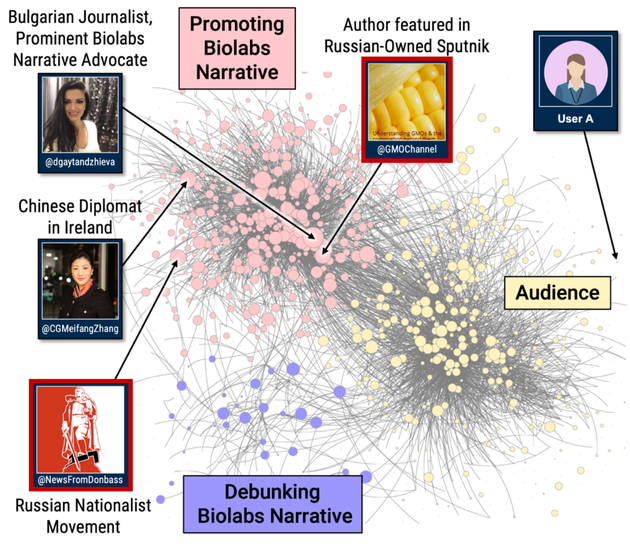
- Large language model embedding puts texts with similar meaning in nearby semantic space
- Clustering in the semantic space reveals the information environment structure
- This semantic structure enables rapid discovery of narrative content without knowing specific keywords
- One hour of human-AI interaction identified 58k of Biolabs narrative tweets out of a 8.5M tweet corpus with 81% precision at 86% recall

AI enables content discovery by semantic meaning, instead of keyword search



Analyst query prompt:
“civilian casualty by Ukraine”

Causal Influence Quantification Reveals Hidden Influencers[1]



Nodes are sized by causal influence and colored by community

Screen name	T	RT	F	Earliest time	Pagerank Centrality	Causal Influence
@dgaytandzhieva	7.3	34.6	49k	01/25/2022	5.6	1,088
@CGMeifangZhang	7.0	16.7	38k	03/08/2022	7.6	822
@User A	8.9	0	800	03/09/2022	0.1	73
@NewsFromDonbass	**3.1**	**0.4**	**3k**	**03/06/2022**	**0.2**	**846**
@GMOChannel	**4.7**	**1.0**	**5k**	**03/07/2022**	**0.2**	**822**

Tweets (T), Retweets (RT), Followers (F)

- ***Causal influence score identifies key influencers that do not stand out in traditional activity and network metrics***
- ***Causal influence models tweet propagation on network and disentangles causal effects from social confounders***

[1] Smith et al. Automatic Detection of Influential Actors in Disinformation Networks, Proceedings of the National Academy of Sciences (PNAS) (2021).

Causal Influence Pathway Quantification Using Network Potential Outcomes

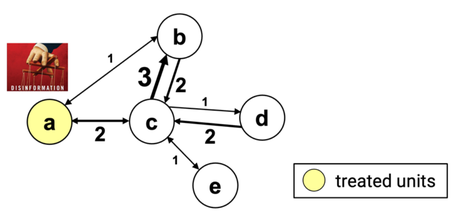
Network Potential Outcomes:[2]

Network potential outcome of unit i *(e.g. number of tweets on IO narrative)*

$$Y_i(\mathbf{Z}, \mathbf{A}_1, \ldots, \mathbf{A}_{n_{\text{hop}}})$$

Influence network

Treatment vector



Causal Influence of Account j over community m:

$$\zeta_j^m = \frac{1}{|\mathcal{C}_m|} \sum_{i \in \mathcal{C}_m} \zeta_{ji}$$

$$\zeta_{ji} \doteq Y_i(\mathbf{Z} = \mathbf{z}_{j+}, \mathbf{A}_1, \ldots, \mathbf{A}_{n_{\text{hop}}}) - Y_i(\mathbf{Z} = \mathbf{z}_{j-}, \mathbf{A}_1, \ldots, \mathbf{A}_{n_{\text{hop}}})$$

Example Path-Specific Causal Estimand – Bridge Influence Estimand:

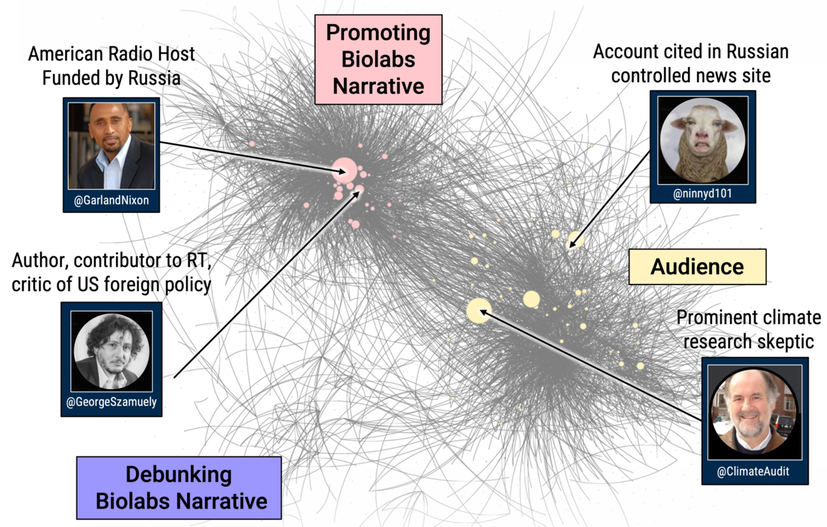
$$\zeta^{\overline{\ell b m}} = \sum_{j \in \mathcal{C}_\ell} \frac{1}{|\mathcal{C}_m|} \sum_{i \in \mathcal{C}_m} \zeta_{ji}^{\overline{\ell b m}}$$

Influence network manipulation counterfactual: Outcome of account i if paths between community ℓ and community m via bridge account b did not exist

$$\zeta_{ji}^{\overline{\ell b m}} \doteq Y_i(\mathbf{Z} = \mathbf{z}_{j+}, \mathbf{A}_1, \ldots, \mathbf{A}_{n_{\text{hop}}}) - Y_i(\mathbf{Z} = \mathbf{z}_{j+}, \mathbf{A}_1^{\overline{\mathcal{C}_\ell \to b}}, \ldots, \mathbf{A}_k^{\overline{\mathcal{C}_\ell \to b, b \to \mathcal{C}_m}}, \ldots, \mathbf{A}_{n_{\text{hop}}}^{\overline{b \to \mathcal{C}_m}})$$

[2] Kao, Causal inference under network interference: A framework for experiments on social networks. Ph.D. Thesis, Harvard University (2017).

Top Bridge Accounts from Biolabs-Promoting to Audience Community

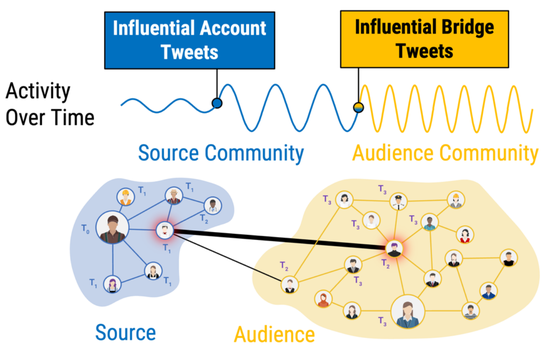


Nodes are sized by causal influence and colored by community

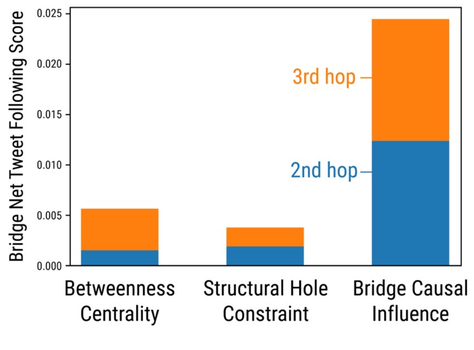
Validation with Net Tweet Following

Bridge Net Tweet Following Metric

How prominent is a bridge in the paths of net tweet following?



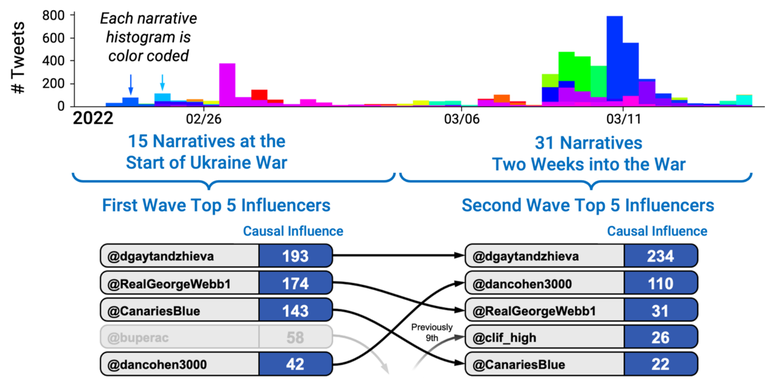
Bridge Net Tweet Following of Top 50 Accounts From Biolab Debunking to Audience Community



Bridge causal influence identifies accounts with stronger evidence of source message amplification in audience community than baseline methods

Results on Predicting Causal Influence and Outcomes

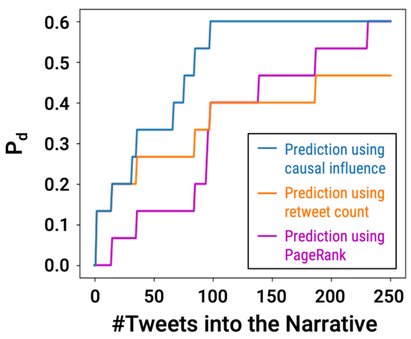
First and Second Waves of Biolabs Narratives



Past causal influence is predictive of future influence

Early Detection of Rising Narratives

Probability of detection (P_d) at constant zero false alarm rate



Participation of past causal influencers is predictive of future rising narratives

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Edward Kao, Yaron Rachlin, Erika Mackin, Andrew Soszynski, Chris Smith, Steven Smith

MIT Lincoln Laboratory, Lexington, MA

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.
This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Air Force.

© 2024 Massachusetts Institute of Technology.
Subject to FAR52.227-11 Patent Rights - Ownership by the contractor (May 2014) Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.