Logical Momentum: Leveraging the BEND Framework to Determine Optimal Patterns and Sequences for Network Influence

Rebecca Marigliano[1,2][0009−0003−8271−9990] and Jacob Shaha[1,2][0000−0002−3963−1462] Dr. Kathleen Carley[1,2][0000−0002−6356−0238]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
[2] CMU Center for Computational Analysis of Social and Organizational Systems (CASOS), Pittsburgh PA 15213

Abstract. The growing importance of social media platforms in shaping the common narrative necessitates a better understanding of how agents interact with and affect the online information environment. Our work investigates how agents within a network build influence over time via sequenced messaging within a social media platform. Utilizing network analysis and the BEND analytical framework, we build a predictive model for influence campaign components, with the goal of identifying the most effective counter-messages to disrupt and defray malicious messaging campaigns online.

Keywords: BEND · Network analysis · Temporal analysis

1 Introduction

Social media constitutes a significant part of the common narrative in information age society [5] [6]. The communication, connection, and influence stemming from social media platforms exert an undeniable influence on the behavior and decisions of individuals, groups, and of society as a whole - often negatively. Social media-derived information has augmented and, in some cases, supplanted the "common narrative" previously the purview of print media and in-person social circles.

Because of this, the perils of misinformation on social media are significant. Ignorant or malicious messages can spread untruths or misleading information with much greater reach than previous narrative mechanisms allowed. The result is widespread vulnerability of social mechanisms, as individual, group, and social understanding can be malformed or controlled by a smaller group of entities than ever before[4].

Identifying and countering the spread of misinformation on social media platforms is a difficult task, and, given the size of the user base, it must be at least partially automated if there is to be any success in real-time moderation and correction. The difficult and interrelated problems of identifying misinformers, detecting misinformation, and counteracting misinformation are all rich areas

of research as social leaders and social media platforms struggle to preserve the utility of this new narrative mechanism [3].

Recent work has shown promise in identifying agents exerting outsize influence on a social media network [7]. Accompanying research can be leveraged to match these agents' influence to potentially malignant or harmful narratives [1]. In combination, these tools can provide an early warning for narrative monitors against parties who may be subverting or eliminating productive dialogue. However, how to react remains an open question. The design and implementation of effective interventions is an open interdisciplinary problem.

Our research aims to answer a small but crucial component of this question: when should information interventions be implemented? Moderators cannot simply counter all potent narrative voices (as identified by the previously mentioned methods), as such would represent censorship and would ultimately suppress genuinely popular and useful viewpoints within the medium. On the other hand, waiting too long to intervene enables harmful narratives to become established and to propagate, potentially removing any ability to effectively counter them. We hypothesize that there exists an ideal moment or range of moments in the evolution of an influencer's communication, where a timely intervention can defuse the specific influence of a harmful narrative without hampering the natural and healthy development of useful narratives.

2 Background

The spread of influence through social networks, both online and offline, has been thoroughly explored in past research [9]. Scholars have demonstrated the utility of quantifying an entity's reach and influence within a network by analyizing the network itself: network metrics derived from graph theory reliable link the potential effect of an agent within a network to that agent's position within the network, and to the structure of the network itself. If we can build a reasonably accurate model of a network, the analysis of that model can provide a quantifiably reasonable measurement of individual influence within the modeled community. We will utilize a similar approach in this paper as we seek to evaluate the "success" of a user's message campaign.

There is much less published work addressing a larger challenge: measuring the cumulative effect over time of many interrelated social media actions. Wei and Carley, among others, have pioneered work to analyze dynamic social networks, examining individual agents' behavior to identify significant deviations or alterations. [10] Such approaches convert static measures of network influence into time-aggregated measures of dynamic activity. Our approach will be significantly simpler, but based on similar reasoning.

An equally daunting challenge is sufficiently quantizing a user's messages to enable quantitative analysis of their effect. A would-be influencer could say or post an endless list of things, each with different content, timing, tone, and scope. All of these communications contribute in subtle and complex ways to the net influence that user builds on the platform, and because of this, the potential

feature space for modeling an influence campaign is intractably massive. To address this dilemma, we utilize the BEND framework developed by Carley and Beskow [2]. BEND provides a compressed quantizing framework for social media actions, mapping message text onto a discrete set of influence "maneuvers" that capture the possible outcomes an influencer may wish to achieve relative to the social network structure and the narratives therein. BEND reduces the nuance and complexity of social media communication to enable categorical analysis. The BEND framework simplifies the study of activities in a multi-agent multicast channel by reducing messages to maneuvers, enabling rigorous and repeatable analysis of content that is otherwise highly variable and subjective. This, in turn, allows us to effectively predict the influence-building effect of a set of BEND maneuvers; in essence, through BEND, we can develop a "small language model" that predicts or dictates the general form of an influence campaign.

3 Methodology

An agent attempting to influence a network and/or a narrative will do so through a campaign of actions: interrelated, sequenced interactions designed to cumulatively produce the desired change in the network/narrative. Using the BEND framework, we reframe an influencer's interactions as a time-sequenced series of BEND maneuvers. We can then evaluate the immediate and cumulative impact of each maneuver to identify the moments of greatest achieved effect relative to the influencer's goal. Such moments represent the "main effect" that moderators would seek to preempt through intervention. We can also try to identify moments in the early sequence when the influencer's campaign is distinct enough to be identified, but has not yet achieved maximum impact. These moments are "main targets" for intervention, as they represent a happy medium between knee-jerk censorship and just-in-time action.

3.1 Data Collection and Preprocessing

Our study utilized Twitter data drawn using search terms focused on highly polarized, binary discussions. In choosing our data, we sought influencers who were oriented to our topic, who demonstrated a reasonable amount of reach or influence (but not a global or super-star level), and who we can assume are actively trying to sway the narrative and/or shape the network. By reducing Tweets based on time and topic, and identifying users who initially used hashtags that would become more widespread, we reason that our resulting targets meet our selection criteria.

3.2 Data Collection and Preprocessing

Our study utilized two datasets: one focused on the Russian-Ukraine conflict and the other on the COVID-19 pandemic.

Russian-Ukraine Conflict Dataset: We collected over 4.5 million Tweets from February 11 to August 30, 2022, using search terms related to the conflict. Influencers were identified based on their engagement with relevant hashtags and demonstrated influence.

COVID-19 Pandemic Dataset: This dataset includes Tweets from 2019 to 2020, focusing on US debates about COVID-19 lockdowns and reopening local businesses. We filtered Tweets to those discussing the "re-open" debate in key states.

To ensure the quality and uniformity of our datasets for analysis, we employed a comprehensive preprocessing pipeline. This included tokenization of tweets to dissect textual content into analyzable elements, standardization of text formats for consistency, and removal of irrelevant terms and known stop words focusing on substantive content. This preprocessing facilitated more effective analysis and enhanced the accuracy of our natural language processing tasks.

3.3 Describing an Influence Campaign

Upon identifying key influencers, we collected their authored Tweets to determine the composition of their influence campaigns. Assuming that all messages sent by these target users were part of their deliberate attempt to build network influence, we used Netanomics' NetMapper software to extract semantic features from the text. We then used Netanomics' ORA network analysis tool to assign BEND labels to each message.

This classification assigns specific BEND maneuvers to each Tweet, such as "boosting" favorable narratives or sowing "dismay" among opposition. A Tweet can contain multiple maneuvers, so we can re-imagine each communication as a variable-length tuple of symbols drawn from the BEND "alphabet." Figure 1 below shows the frequency of each BEND maneuver within the dataset.

We also considered the temporal clustering of Tweets to describe the influence campaign accurately. If the time elapsed between two consecutive Tweets from the same user was too large, we treated the second Tweet as the start of a new influence campaign. This approach ensures that our analysis reflects coherent sequences of influence attempts rather than sporadic activity. This method was applied consistently to maintain the integrity of our influence campaign analysis.

3.4 Measuring an Influence Campaign

As previously described, we use network metrics to measure the "success" of our targets' respective influence campaigns. Crucially, we avoid a measurement scheme that would incorrectly attribute any network metric to any specific message, as that grossly oversimplifies the phenomenon we are studying. Instead, we build our metrics with the goal of identifying *cumulative impact* over the course of a campaign. Toward that end, we focus on the *change* in metrics surrounding each tweet, rather than the value of the measurement.

Our constructed network is a bimodal directed graph, built of connections between users and tweets based on the observable propagation of messages within

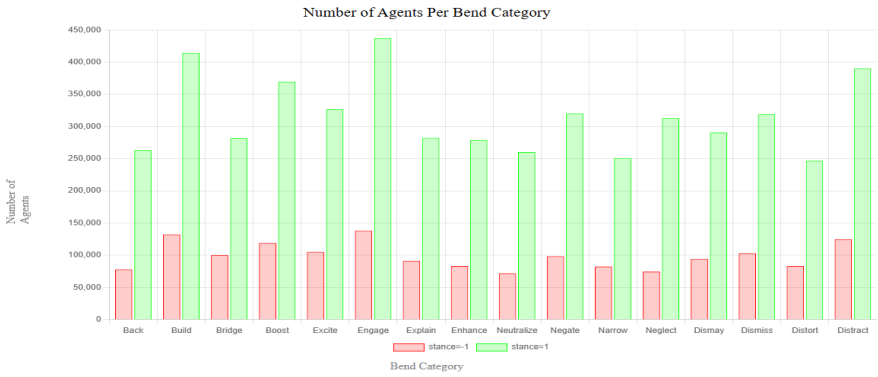

Fig. 1. BEND Maneuver frequency within the dataset. Green bars are Pro-Ukraine occurrences; red bars are Pro-Russia.

the dataset. We use two different measures of centrality when measuring our targets' influence within this network: the *in-degree centrality*, or number of connections pointing *to* the user; and the *out-degree centrality*, or the number of connections pointing *away* from the user. These metrics, considered together, provide different measures of a node's importance in the graph. We also consider each user's *ego network*, a subsection of the graph that captures the sub-network directly visible to that user. We measure the density of this network as a way of tracking how immediately impactful the user is on the graph structure, as any attempt to create or dissolve social groupings would be most obvious in the user's ego network. As an additional metric, we record the engagement each Tweet receives in the form of retweets and/or quotes. This measure is not as effective at conveying cumulative impact, but is still a useful measurement for identifying particularly successful phases of a campaign.

To derive our metrics, iterating through our targets, we collected all their authored tweets. Then, for each tweet, we captured the time stamp t and constructed two networks: one from all the tweets up to t, and one from all the tweets up to time $t + \delta$. We then measured the in- and out-degree centrality of these two networks and took the difference between them. That difference was assigned to the pivot tweet as its metric score. We reason that the contents of the tweet at time t were, at least partially, responsible for subsequent changes in the network. By comparing the network at the time of the message to the network some time δ afterward, we quantify and capture that induced change.

Examples of the resulting metrics are shown in Figure 2. We plot derived values in tweet-order, showing the change in the measured values over the run of the sequence. Effectively, our influence campaign metrics display the first derivative of the actual network characteristic, with respect to the target influencer's messages.

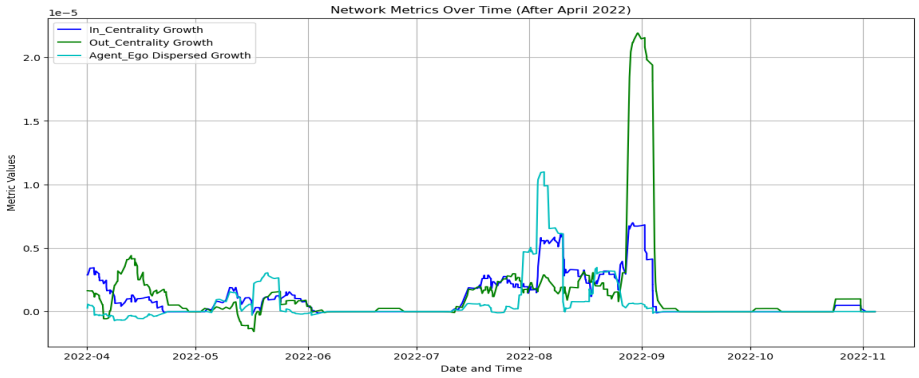

Fig. 2. Example user metrics over time.

3.5 Modeling and Prediction

Having defined our features and our metrics, we next attempted to model the system at play. We chose a random forest regression model, since our array of 16 binary features are likely to have non-linear effects on the final metric values. Our model had unlimited tree depth and contained 100 different estimating trees. To train our model, we excluded one user's tweets from the data, and cross-validated on the remaining tweets to identify the best model parameters. We compared models using the Friedman-corrected mean squared error as our scoring function.

We then tested the trained model against the excluded user. We repeated this process for each user, predicting that user's influence metrics from all other users' tweets, and took the average of all model coefficient of determination R^2 scores to indicate the overall success of our method. We repeated this process for each metric separately, as well as training a multivariate-outcome model against all metrics simultaneously.

4 Analysis and Follow-on

Our model results are given in table 1. As mean-squared error is a positive value, the listed median and standard deviation represent a one-sided distribution. We give the median to provide resilience against outliers, as many models had wildly inaccurate predictions early in the sequence before rapidly converging toward the actual series values. The model score value R^2 listed has a maximum value of 1.0, indicating a perfectly predictive model; negative scores indicate arbitrarily incorrect predictions. [8] Figures 3 and 4 show examples of a best-model prediction versus a single target user. The solid line is the difference between the single-metric predictive model and the measured value; the dotted line is the difference between that metric in the multivariate model and the actual value.

Table 1. Summary of Metrics Across Different Parameters.

Metric	Measure	Mean value	Median value	Std. Deviation
Tweet engagement (x_1)	*Raw value*	200.9k	4.9k	451.7k
	Model avg	-1.626	-1.215	1.144
	Best model	0.015	0.02	0.085
In-degree cent. (x_2)	*Raw value*	4.01e-9	0.05e-9	15.36e-9
	Model avg	0.85	0.86	0.025
	Best model	0.976	0.977	0.008
Out-degree cent. (x_3)	*Raw value*	1.492e-9	0.277e-9	3.305e-9
	Model avg	0.186	0.194	0.518
	Best model	0.983	0.99	0.012
Ego-net density (x_4)	*Raw value*	2.083e-3	0.479e-3	3.281e-3
	Model avg	0.2	0.246	0.146
	Best model	0.799	0.77	0.072
Results vector ($\mathbf{x}$)	*Raw value*	42.4k	2.4k	84.2k
	Model avg	-5.147	-0.777	12.13
	Best model	0.334	0.326	0.058

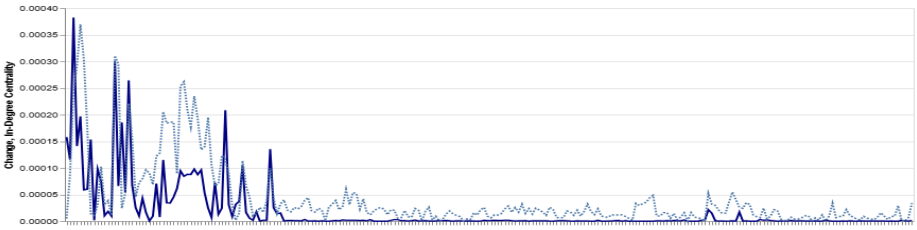

Fig. 3. Prediction error for in-degree centrality, for a sample user, for single-variable (solid) and multi-variable (dotted) models.

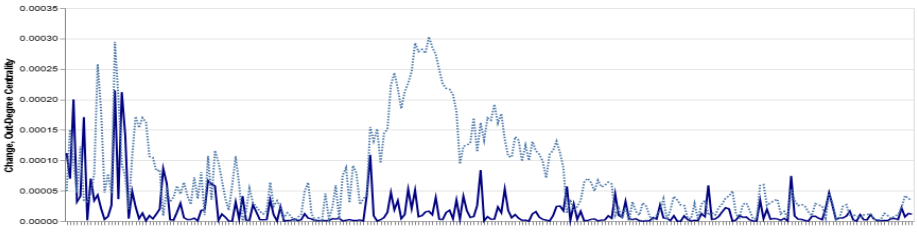

Fig. 4. Prediction error for out-degree centrality, for a sample user, for single-variable (solid) and multi-variable (dotted) models.

Our results varied widely across our influence metrics. Notably, our poorest performance was in metric 1, Tweet Engagement. Crucially, this metric was the only non-comparative measurement (i.e. it measured directly from a tweet at time t rather than capturing an effect accumulated over time interval $[t, t + \delta]$). We hypothesize that this "instant response" characteristic made our inherently autocorrelative approach ill-suited to predict that metric.

Our models were similarly underwhelming for metric 4, Ego-net Density. This warrants further investigation, as there may be systemic issues with how we generate and evaluate ego networks from our truncated Tweet corpus. However, it may also be the case that agents have significantly *less* direct control over their ego networks than we might have expected. In other words, an agent's influence attempts may more heavily impact the larger network, which then shapes the agent's ego network, an inversion of our working theory.

We saw success in predicting change in network centrality based on BEND maneuvers. Our best models averaged scores of 0.976 and 0.983 for in- and out-degree centrality respectively, very near the highest possible score of 1.0. As seen in the example figures, some perturbations persist throughout the sequence, indicating instances where the model failed to correctly predict sequence behavior. Crucially, though, the model correctly predicted the *direction* of influence change for 97.5% of data points for centrality metrics, and 98.4% for egonet density. So, while the model may often overestimate the *size* of an effect, it is highly accurate at predicting whether a sequence will result in a *gain* or a *loss* of influence.

Our research has demonstrated a promising avenue for decomposing and better understanding online influence campaigns, a vital first step in effectively countering malicious influencers. To expand on our work, we plan to incorporate similarly processed campaigns from other cultural, topical, and temporal settings, to investigate how resilient our modeling approach is to changes in communication style or topicality. We also plan to investigate whether other, more effective success metrics can improve our accuracy, by folding our directed bimodal graph into an undirected unimodal graph. While this discards some of the network's nuance, it facilitates the use of many more network metric algorithms. Finally, we plan to investigate additional features that may better inform our model's predictive power, including adding a time-delay feature to each tweet measuring the elapsed time between messages.

Applying our research to the root problem - intervening to derail malicious influence campaigns - is now in reach. With the trained model, a user can contruct a set of all possible maneuver combinations (2^{16} possibilities) and have the model evaluate them. The user can then select the best-scoring combination to increase their influence, or the worst-scoring combination to reduce it.

However, such application still faces several hurdles. Chief among these is the difficulty of conducting a controlled experiment that accurately models the widespread and multi-layered nature of online social media platforms and attaining influence therein. We expect very limited capability to use our research (as presently constituted) to craft and deploy media interventions, due to the incomplete nature of our work and the ethical restrictions inherent to such actions.

Finally, even if we can estimate which maneuver combination will best reduce a target's influence, it is unclear if that maneuver combination will have the same effect when produced by a third party as opposed to the agent themselves.

We plan to combine our research with ongoing efforts to develop robust and complex agent-level simulations of social networks. Our models can be tested in such an environment by dictating an agent's strategy as they attempt to affect the network and the narrative. And, we can test the intervention strategy efficacy in this setting to investigate whether third-party influence "attacks" follow the same predictive patterns.

Acknowledgments. We are indebted to the many researchers who have built a robust and thorough foundation in social network research. We especially recognize the contributions of Carnegie Mellon's CASOS and IDeaS centers. These contributions include the BEND framework which, regardless of its shortcomings, provides a useful and standardized methodology to quantitatively consider issues like this one. Our ability to understand and explore our network data was vastly enhanced by Netanomics ORA and NetMapper software.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Benigni, M., Joseph, K., Carley, K.: Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining, chap. Bot-ivistm: Assessing Information Manipulation in Social Media Using Network Analytics. Springer, Cham (2019), https://doi.org/10.1007/978-3-319-94105-9_2
2. Carley, K.: Social cybersecurity: an emerging science. Comput Math Organ Theory **26**, 365–381 (2020)
3. Carley, K.: Disinformation: A Multi-National, Whole of Society Perspective, chap. A Political Disinfodemic. Springer, Cham (2022), https://doi.org/10.1007/978-3-030-94825-2_1
4. Fujiwara, T., Müller, K., Schwarz, C.: The effect of social media on elections: Evidence from the united states. Journal of the European Economic Association (2023), https://doi.org/10.1093/jeea/jvad058
5. Holmstrom, M.: The narrative and social media. Defence Strategic Communications **1**, 118–132 (2015)
6. Muhammed T, S., Mathew, S.K.: The disaster of misinformation: a review of research in social media. International journal of data science and analytics **13**(4), 271–285 (2022)
7. Ng, L.H.X., Carley, K.M.: Online coordination: Methods and comparative case studies of coordinated groups across four events in the united states. In: Proceedings of the 14th ACM Web Science Conference 2022 (2022)
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

9. Wasserman, S., Faust, K.: Social network analysis. Cambridge University Press (1994)
10. Wei, W., Carley, K.M.: Measuring temporal patterns in dynamic social networks. ACM Trans. Knowl. Discov. Data **10**(1) (jul 2015). https://doi.org/10.1145/2749465, https://doi.org/10.1145/2749465