Modeling Conspirators in Virtual, Text-Based Communities

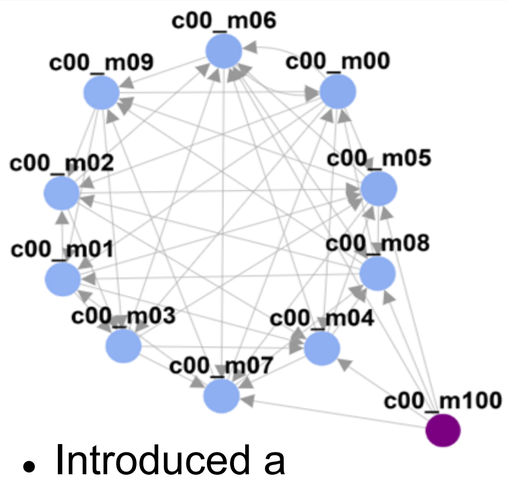
Problem Statement and Goals

In response to the 2019 recommendations from the U.S. National Academies of Sciences, Engineering, and Medicine to the U.S. Intelligence Community on research into online influence (NASEM 2019), we focus on understanding how internal conspirators influence the dialogue and, therefore, members of virtual, text-based communities.

Research Question:

What is the effect of changing the target type from the most influential to the most influenced member for conspirators using the always agree and always disagree strategies?

Approach



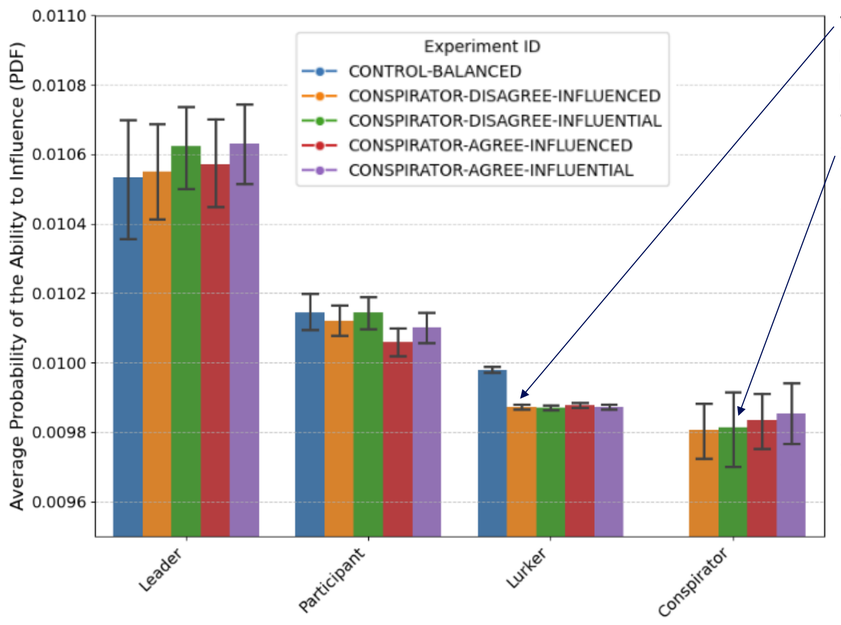
- Introduced a Conspirator (c00_m100) to the community
- Used a multi-agent model of a text-based virtual community that modeled each member using a separately prompted LLM
- Assumed the 90:9:1 model of member-level activity
- Each member has an anchoring bias and a tailored prompt that informs its dialogue

Conspirator Parameters that Can Be Manipulated

Target Type	Dialogue Strategy	Enter Conversation Turn
MOST INFLUENTIAL	ALWAYS AGREE	Conversation turn in which the Conspirator member begins dialogue with its 1-hop neighbors. We chose 3 out of 20 conversation turns in an experiment.
MOST INFLUENCED	ALWAYS DISAGREE	

Results

Distribution of Members' Ability to Influence

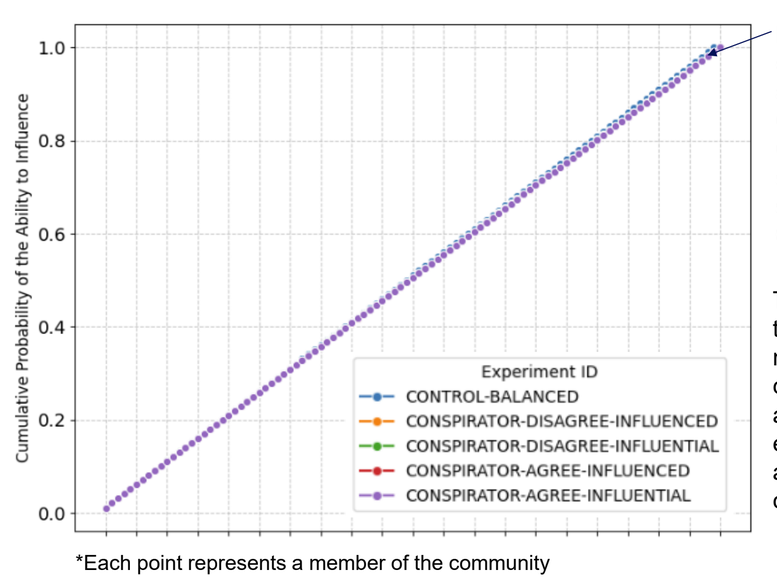


The introduction of a Conspirator member reduced the probability of a Lurker's ability to be influential by approximately 1.1%.

The probability of a Conspirator's ability to be influential was 1% less than other members on average.

These interim results prompt the following questions for future research:

1. Why do Lurkers' average probability of being influential decline with the introduction of a Conspirator?
2. Why is the Conspirator less influential than the Leader, average Participant, and average Lurker?



*Each point represents a member of the community

Non-parametric tests comparing the CDF of the ability to influence between the CONTROL and each Conspirator experiment showed no statistically significant difference at the 0.05 level.

Therefore, changing the target type from most influential to most influenced for conspirators using the agree and disagree strategies had no effect on the distribution of the ability to influence in the community.

NOTE: We initialized the community members with an equal number of members with CON, NEU, and PRO stances. However, we found that the GPT-4o LLM was unable to maintain a neutral stance during the dialogue, which could introduce a bias into these results. This will require more detailed experiments to understand.

Future Work

Once we understand how conspirators work in virtual, text-based online communities, we will study how to detect an ongoing influence operation.

References:

National Academies of Sciences, Engineering, and Medicine. 2019. "*A Decadal Survey of the Social and Behavioral Sciences: A Research Agenda for Advancing Intelligence Analysis*." Washington, DC: The National Academies Press. https://doi.org/10.17226/25335.

Pinto-Coelho, Ciro, Straub, Alexander, Waller, Addison, and Cukier, Michel. "*Dialogue vs. Topography: Assessing Influence Metrics in an Agent-based Model of Online Communities*." Manuscript submitted for publication.

Acknowledgements: Thank you to Ciro Pinto-Coelho and Dr. Michel Cukier for advising and support. Thank you also to Amrit Magesh and Aaron Chen for collaborating on the model and analysis techniques.

Sofia Douglass
University of Maryland, College Park

Modeling Conspirators in Virtual, Text-Based Communities

Abstract

In response to the 2019 recommendations from the U.S. National Academies of Sciences, Engineering, and Medicine to the U.S. Intelligence Community on research into online influence (NASEM 2019), we focus on understanding how internal conspirators influence the dialogue and, therefore, members of virtual, text-based communities. To accomplish this, we use the Aletheia Third Generation (A3G) multi-agent model, which consists of large language models (LLMs) used to simulate members of a virtual community talking. The framework consists of two conversation models: public (*e.g.*, blog discussions) and private (*e.g.*, SMS or e-mail). This research focuses on the private model.

Communities in the A3G model follow the 90:9:1 model of member activity, with 90 Lurker members, 9 Participant members, and 1 Leader member. Each member has 1-hop neighbors, which are the other members in the community with whom they can communicate. Leaders are most active and have the most 1-hop neighbors. Participants are less active with fewer 1-hop neighbors, and Lurkers are the least engaged and have the fewest 1-hop neighbors. This research introduces a Conspirator member who is about as active as a Leader member. The Conspirator member takes its target member and that member's 1-hop neighbors as its 1-hop neighbors. The message a member sends to each of its 1-hop neighbors affects that neighbor's stance on the topic. A member's ability to change the stance of one of its 1-hop neighbors determines its influence score. These influence scores are used to examine the distribution of members' ability to influence in a community.

A Conspirator member enters a community after its members have begun their discussions. We initialize it with a target type (Most Influential, Most Influenced, or Random) and a strategy (Always Agree, Always Disagree, Initially Agree, Initially Disagree, or Ignore). This poster explores the Most Influential and Most Influenced target types with the Always Agree and Always Disagree strategies. We found that changing the target type from Most Influential to Most Influenced for Conspirators using the Always Agree and Always Disagree strategies had no statistically significant effect on the distribution of the ability to influence in the community. Our interim results also suggest the need to conduct future work to address why Lurkers' average probability of being influential declined with the introduction of a Conspirator and why the Conspirator was less influential than the Leader, average Participant, and average Lurker.

By understanding how changes to the Conspirator's target member and strategy affect the distribution of the ability to influence, we can work towards answering the question we posed in response to the social and behavioral sciences research program recommendations that the U.S. National Academies of Sciences, Engineering, and Medicine presented to the U.S. Intelligence Community regarding research into online influence (NASEM 2019): How do we detect an ongoing influence operation in a virtual text-based community?

References:
National Academies of Sciences, Engineering, and Medicine. 2019. "*A Decadal Survey of the Social and Behavioral Sciences: A Research Agenda for Advancing Intelligence Analysis*." Washington, DC: The National Academies Press. https://doi.org/10.17226/25335.

Sofia Douglass
University of Maryland, College Park

UNIVERSITY OF MARYLAND