Examining Influence Dynamics in Online Communities Through a Public Conversation Model

Addison Waller
University of Maryland

Abstract

Influence is a widely studied phenomenon that requires a deeper understanding of its impact on human communities. With the creation and heavy use of social networks, the ability to influence a large group of people has become easier than ever. Influence can be good for spreading important news but also can be harnessed negatively to spread misinformation. Due to this, it is critical to gain an understanding of how influence spreads throughout an online community and what measures can be taken to prevent it. In this paper we create an agent-based online community framework called Aletheia - Third Generation (A3G). With this framework we are able to represent text-based online communities with the intent of understanding how information and influence spread. This paper focuses specifically on the development and creation of the part of A3G that is structured like a blogging site, called the public conversation model. We discuss how this model can be used to answer a variety of questions about the spread of influence and detection and mitigation of influence operations.

Introduction

Attempts to influence people over a form of media has been around for decades. When done by a company, religion, or government to further a goal or spread an ideal this is referred to as propaganda. With the move to online communities and social networks, the amount of information shared around the world has vastly increased. This increase has allowed for an increase in coordinated attempts to change public opinion with groups of online bots, often done by state actors, called influence operations. Johnson et al. (2020) discuss the direct impact foreign influence operations were able to have on the 2020 election (Johnson, et al., 2020). Even though the impact of influence and influence operations have become a realized threat, there is a call from the academic community to gain a specific understanding of how influence spreads through an online community and what metrics can be used to measure it. This was specifically stated in the 2019 National Academies of Science, Engineering, and Medicine study when they pointed out the need for metrics for influence that go beyond “vanity” metrics (e.g., “likes”), methods to measure the effectiveness of an influence campaign, and methods to identify “at-risk” groups and strengthen their resilience to influence (National Academies, 2019).

These findings were our motivation to create an agent-based model of text-based online communities to study influence. To properly create this model there had to be an understanding that influence operations can occur in a wide variety of places. This model uses pretrained large language models (pLLMS) to enable the generation and cognitive processing of messaging between members of a text-based online community. Our conversation model is called Alethteia - Third Generation (A3G) which uses third generation pLLMs like GPT 3.5 to act as a member. When focusing on text-based online communities, as we do in this study, there are usually two categories of communication platforms. The first is private communication platforms, these are similar to text messaging and email. The second are public communication platforms which resemble the structure of blogs and forums, like Reddit. Reflecting these, A3G contains two different types of online communities, the private conversation model and the public conversation model. The private conversation model is built to reflect emails where the public conversation model is structured to reflect a community that is able to create a variety of posts and comment on them, like a blog. This paper will focus on the design of the public conversation model by covering sections about related work, the background of the A3G, the Public Conversation Model, and Future Work.

Related Work

There are two main research disciplines that are important to give the public conversation model a good background. They are online communities and agent-based modeling. First it is important to define an online community. Kraut states that an online community is "…any virtual space where people come together" with only two requirements of "…ongoing interactions among people over time…" and mediated by technology (Kraut, 2012). A major piece of creating an online community is how to structure the members in them. There are a variety of ways to create and structure these members so that they most resemble a real community. One classification system is the 90:9:1 rule. This is where 90% of members are lurkers, 9% of members are participants, and 1% are leaders (Nonnecke, 2000; Nielsen, 2006). It has been shown that most members in an online community do not interact but simply take in the information, this is where the 90% lurkers come in. Another member classification system is "Reader-to-Leader" which has four member categories from least to most active (Preece, 2009).

The second research discipline that is important to understanding the public conversation model is agent-based modeling. Agent-based modeling is helpful to use when building an online community because it allows the focus to be on the activities of the members as opposed to the overall structure of the community. Epstein shows that larger outcomes, like forming a group, are able to come from the smaller characteristics of an agent-based model. These characteristics are heterogeneity of agents, autonomy of agents, a well-defined space in which the agents interact, and agents actions are based solely on local information (Epstein, 2006). There have been multiple types of agent-based community structures created over the years that rely on probabilistic models of the interactions but the newest research focuses on the use of pLLMs as agents. It has been demonstrated that the GPT-2 pLLM (Radford, 2019) can be fine tuned to instill a persona (Boyd, 2020) and that a collection of these models can act as a community of people discussing a topic (Betz, 2021). Along with this, Touvron was able to prove that using prompt engineering a convincer pLLM can use elements of knowledge and trust to influence a skeptic pLLM (Touvron, 2023). These studies allow us to confidently use pLLMs as our agents because they can represent humans in an online community.

Background

A3G is a framework of pLLMs that can be used to simulate members talking in a variety of social network situations. The public conversation model is a conversation turn-based model meaning that each member goes once during a conversation turn until the experiment is over. Additionally, a member can be considered offline or online. When a member goes offline during a conversation turn it is unable to comment during that turn. The amount of time a member spends offline or online can be randomized or fixed. The goal of offline/online is to show the appearance of someone logging off a computer or taking time away from the social network. The type of member also has an impact on how much a member posts or comments. After the completion of all conversation-turns, A3G produces an online community that mimics real online communities. The specifics of member and community implementation will be discussed in the subsections below.

Once the A3G framework has been fully developed the goal is to be able to answer four broad research questions:

- *RQ1*: How do we define and quantify influence in a text-based online community and what level of agreement exists between this method and existing methods?
- *RQ2*: Can we define characteristics of an online community that would put it more “at-risk” to an influence operation?
- *RQ3*: Can we detect either the onset or existence of an influence operation in a text-based online community?
- *RQ4*: Once detected, can we define steps to mitigate the effects of an influence operation?

These research questions will eventually allow for the creation of proper experiments to test influence in an online environment as well as influence operation detection and mitigation.

Communities Implementation

In A3G a community is a group of members that interacts with one another. In the Private Conversation Model there can be a variety of communities. The members in each community mainly interact with each other but have the ability to occasionally interact with another community. The Public Conversation Model only has one community per experiment with each post in the community focusing solely on one topic from different viewpoints.

Member Implementation

Members are pLLMs that are given specific parameters in order to create a community of unique individuals. Each member is made up of five main parameters: Cognitive Bias, Persona, Perspective, Stance on a Given Topic, and Member Type. Cognitive bias has to do with how the process of thinking and making decisions is affected (Haselton, 2015). This occurs naturally in humans when they are presented with a lot of information. In this work we use four types of cognitive biases.

- *Anchoring Bias:* The member takes to the first piece of information it encounters and uses that to guide its behavior
- *Confirmation Bias – Opposite (CB-O)*: Member is attracted to information with opinions that oppose theirs and so limits its communication to 1-hop neighbors that have a different stance on the topic of discussion
- *Confirmation Bias – Same (CB-S)*: Member is attracted to information with opinions that are the same as theirs and so limits its communication to 1-hop neighbors that have the same stance on the topic of discussion
- *Conformity Bias:* Member tends to take cues for proper behavior from the actions of others rather than exercise its own independent judgment – selects a stance that is the most prominent stance among its 1-hop neighbors at each conversation-turn
- *Status Quo Bias:* Member favors keeping things as they are and so never changes its stance

Each of these cognitive biases determines how the member interacts with the information it encounters in the community. It is critical to craft our pLLMs to impersonate a real person so our crafted online community will represent real online communities. In a similar manner, a persona represents the identity of the member. Through prompt engineering the member is provided with details about who it needs to be and how it should interact with information it encounters. An example of a prompt to create a persona would be:

> You are not an AI language model. You are the leader of an online community discussing whether to legalize drugs in the USA. You are rather close-minded. You are only interested in ideas about whether to legalize drugs in the USA that are the same as yours. You are unafraid of politely but forcefully convincing other members in the community of your point of view. Your responses are based on your background beliefs and thoughts.

The third main parameter that makes up a member is perspective. Perspective is the opinion a member holds about the overall topic being discussed in the online community. Perspective is initialized based on input parameters and then updated at each conversation-turn using text generated by the pLLM at that turn. Stance, the fourth member parameter, is determined by using the member's perspective. A member's perspective is fed into a separately prompted pLLM that determines the message's stance. A stance is represented by PRO, CON, or NEU with the added variation of HIGH, MID, or LOW depending on the strength of the stance. An example stance would be PRO-HIGH if the member has a highly favorable opinion of the current topic.

The final major parameter that goes into creating a member is the member type. There are three different member types a member can be given: leader, participant, or lurker (Preece, 2009). A leader is a member who is statistically much more likely to make community contributions as well as usually connected to a higher number of other members. There are usually a small number of leaders per community. A participant is far more common in a community. They usually make contributions around half the time but have fewer connections than a leader. A lurker is the most common member type in a community but interacts with it the least by making contributions only very sporadically (Nonnecke, 2000; Nielsen, 2006).

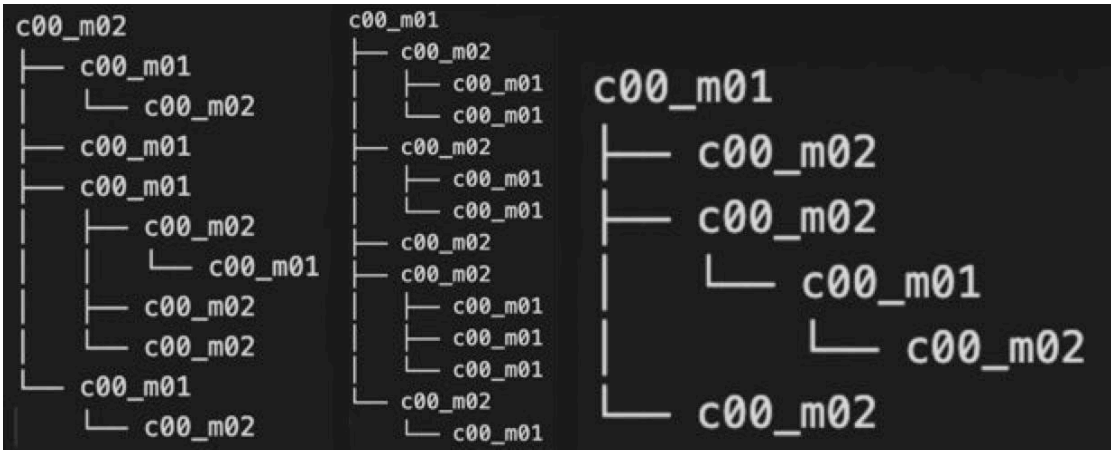
Public Conversation Model

Public Conversation Model Design

The goal of the public conversation is to create an online community that has a blog structure. The main inspiration for the public conversation model came from Reddit subreddits. A Reddit subreddit is a community on Reddit that is created to discuss a specific topic. To recreate this, each public conversation model run only has one community that discusses a specific provided topic. While the topic stays consistent the members' opinions on the topic can vary greatly. During each conversation turn all online members will have the opportunity to either make a post or a comment on a post or another comment. A post in a community for the public conversation model is a statement made by the member about the overall topic. A comment is a statement made in regards to a specific post or comment under a post. To ensure the experiment has diversity there will be no set conversation turns in an experiment rather a maximum depth of comments on a single post. Once that depth is hit on any of the posts the experiment will end. Along with this there will be a set number of posts that can be made in one community.

Posting and commenting

As mentioned above a post is a standalone statement in a community discussing a specific topic where a comment is made with the context of the overall topic and the statement, whether that is a post or another comment, the member decided to comment on. *Figure 1* shows an example of the tree structure used for communities where each tree represents a post. In *Figure 1* there are three posts of different sizes where c00_m02 represents member 2 in community 0. In the first tree it can be seen that c00_m02 is the agent who made the post and all the members below it are the ones who commented. When a member gets an option to post or make a comment the statistical chance of them doing so is based on their member type, stance, and cognitive bias compared to the statement's stance.



```
c00_m02
├── c00_m01
│   └── c00_m02
├── c00_m01
├── c00_m01
│   ├── c00_m02
│   │   └── c00_m01
│   ├── c00_m02
│   └── c00_m02
└── c00_m01
    └── c00_m02
```

```
c00_m01
├── c00_m02
│   ├── c00_m01
│   └── c00_m01
├── c00_m02
│   ├── c00_m01
│   └── c00_m01
├── c00_m02
├── c00_m02
│   ├── c00_m01
│   ├── c00_m01
│   └── c00_m01
└── c00_m02
    └── c00_m01
```

```
c00_m01
├── c00_m02
├── c00_m02
│   └── c00_m01
│       └── c00_m02
└── c00_m02
```

Figure 1: This gives an example tree structure of what a community with 3 posts of different sizes looks like. c00_m02 structure is the member 2 in community 0 is the member who created the post. Each of these members represents a post they made.

When deciding whether to make a post or a comment each member will go through a specific process. First, the member will decide whether or not to make a post. If they choose to create a post then that is all they will do for that conversation-turn. If they decide not to post they will then be given the opportunity to make a comment on another post. If they also decide not to comment, their conversation-turn will simply end. If they do decide to make a comment they will read through each post in the community and decide which one to comment on based on the member's current stance and cognitive bias as well as the post's current stance. The specifics of which cognitive bias will select which type of stance can be seen in *Table 1*. If no posts fit with the members persona then they do not comment and end their conversation-turn. If the member is able to locate a post then they will either comment directly on the post or read the comments under the post and select one to comment on. This process of reading the comments and selecting one to comment on is done the same way as selecting a post.

The chances of a member deciding to post or comment is based on their member type. Leaders have the highest chance of making a post or a comment with a 95% chance. This is due to the fact that they are the most vocal in a community. Participants are still active in a community but to a much lesser degree than the leaders so they have a 55% chance of making a post or comment. Finally, lurkers have the lowest chance of making a comment or post with a 1% chance.

Cognitive Bias	Stance they prefer
Anchoring Bias	The first post they read will become their “anchor”. After that they will comment on whatever statement closely matches their stance.
Confirmation Bias – Same	Only comment on whatever statement closely matches their stance
Confirmation Bias - Difference	Only comment on whatever statement is the most different from matches their stance
Conformity Bias	Comments on the stance that is most prevalent in the comments they read
Status Quo Bias	Comment on the statement that is the same as its original stance. Tries to never change its stance

Table 1: What type of post/comment stance each cognitive bias prefers to comment on

Influence

Influence is convincing or persuading someone to do an action or embrace a belief they normally would not. Influence was slightly more difficult to calculate than it is in the private conversation model. As a result, influence is calculated based on who a member comments on. Since the member reads all the posts or comments before deciding to comment, the statement they responded to clearly caught their attention in relation to their stance. If a member decides to directly comment on a post that post will be the only statement used in the calculation of influence metrics for that member at that conversation-turn. When a member decides to comment on a comment instead of directly on a post calculating influence becomes more complex. *Figure 2* shows examples of these influence structures. By adding the influence of the post to the influence calculations, along with the comment, the impact of the post and the comment that was selected are properly represented.

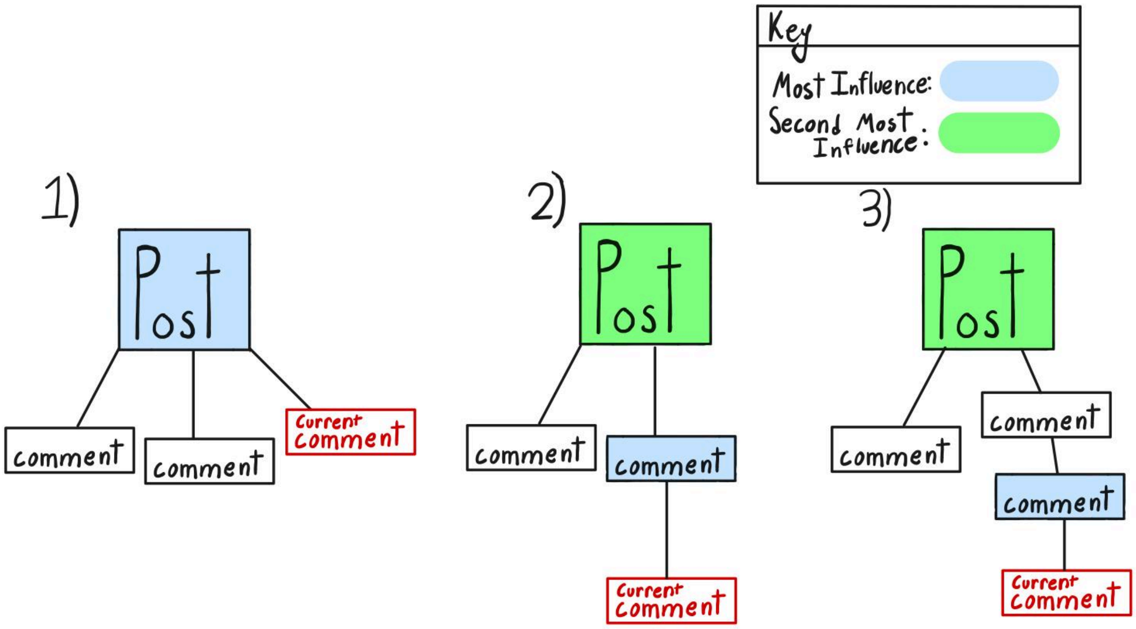


Figure 2: Three different post tree structures are displayed in this Figure. The blue highlight reflects the statement that is the most influential to the current comment and the green highlight reflects the second most influential.

Metrics

There are two major metrics that are calculated at each conversation turn. The first metric focuses on who were the most influential members for this turn based on each member's influence score. The second metric is used to determine the community's polarization. This is measured by computing the ratio of occurrence versus number of members for each stance in the community. All of these data from each conversation turn is later used to output a large range of data that can be analyzed. These data include most influential members from each turn, influence scores for each member at each turn, member stance at each turn, the full conversation, and more.

Current State of Public Conversation Model

Starting Point

The public conversation model is a framework that has been in development for around 2 years. During this time it has gone through many iterations. When I took over the public model about a year ago I was tasked with integrating it into the current Aletheia structure but to keep the implementation of the public conversation model the same. Originally the public conversation model had a thread-based structure. This means that each member was run as a thread and when they were online they would lock on shared post objects to add their comments. While this method was working there were a range of reasons why we decided to change the implementation to a turn-based system instead. Due to the fact that we were already integrating the public conversation model into Aletheia it is logical to make the public and private conversation models have a similar structure. This way both models would produce the same type of results. These same types of results allow for the public and private conversation models to be easily compared. Along with this, the thread-based structure is much more complex in function and harder to maintain than the turn-based structure. By changing to the turn-based structure we eliminate the possibility of vulnerabilities like data races. Finally, no metrics had been implemented into the thread-based structure of the public conversation model. Converting to the same structure as the private conversation model meant that we were able to add its metrics to work with the public conversation model with little effort.

Accomplishments

In the year I have been working on the public conversation model, I have been able to almost fully incorporate the model into Aletheia and update it to a turn-based conversation structure. To update this I used the foundation code from the private conversation model to base the offline/online system for the public model. Then I updated the communication functions, which are responsible for creating posts and comments each turn. This entailed quite a few changes, but the biggest one was updating the commenting functionality so it selects a post or comment to comment based on the member's cognitive bias instead of randomly. Afterwards, the metrics from the private conversation model were converted to work for the public conversation model's needs. Outside of these main functions which allow the public conversation model to run, the member and community classes also needed to be updated in order to fit in with the new non-thread based design.

Next Steps

AG3 is a long term research project that has many moving pieces and will continue for longer than a single year. Because of this, there still remains a few things to be done with the public conversation model to get it up and running. First off, the final piece of the framework is coding the influence and perspective calculation. The subsection above, Influence, discusses how we plan to calculate influence. Although we know our intended way to calculate influence, the implementation itself is more difficult and will require a couple additional weeks. Once implementation of the framework has been completed it will need to be tested. Testing will include running a diverse set of experiments to search for persistent errors. After all major errors have been addressed we can begin using the framework to answer our posed research questions.

Future Work

A clear direction for the future of the public conversation model is to use it to answer the four research questions posed in the Background section. Understanding how influence spreads throughout a community is the main reason for creating this framework. Using this framework to explore the research questions will bring us closer to gaining a deeper understanding of how influence spreads and preventing negative spread. In order to explore research questions 3 and 4 the public conversation model must implement something that has recently been implemented into the private model: conspirators. Conspirators are a malicious member type which attempts to place misinformation into a community or cause mayhem. By adding these conspirators we will begin to be able to study influence operations and understand how to detect and mitigate them.

Conclusion

As discussed throughout this paper, understanding how influence spreads through a community is critical to being able to detect and prevent malicious influences. With A3G we will be able to study how influence spreads by allowing us to control what information is introduced into a community and observe how it propagates. The public conversation model will specifically allow for studies of blog structured networks. Future work will have the model able to emulate more realistic online communities in order to understand the spread of influence and develop software to stop and prevent it.

References

Betz, Gregor. “Natural-Language Multi-Agent Simulations of Argumentative Opinion Dynamics.” Journal of Artificial Societies and Social Simulation 25, no. 1 (January 2022): 2. https://doi.org/10.18564/jasss.4725.

Boyd, Alex, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. “Large Scale Multi-Actor Generative Dialog Modeling.” arXiv, May 2020. https://doi.org/10.48550/arXiv.2005.06114.

Epstein, Joshua M. 2006. “Chapter 1: Agent-Based Computational Models and Generative Social Science.” In Generative Social Science: Studies in Agent-Based Computational Modeling, STU-Student edition. Princeton, NJ, USA: Princeton University Press. http://www.jstor.org/stable/j.ctt7rxj1.

Haselton, Martie G., Daniel Nettle, and Damian R. Murray. “The Evolution of Cognitive Bias.” In The Handbook of Evolutionary Psychology, 1–20. John Wiley & Sons, Ltd, 2015. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119125563.evpsych241.

Johnson, C. et al., 2020. Foreign Interference in the 2020 Election: Tools for Detecting Online Election Interference, RAND Corporation. United States. Retrieved from https://policycommons.net/artifacts/4836708/foreign-interference-in-the-2020-election/5673421/ on 24 Apr 2024. CID: 20.500.12592/4wmj61.

Kraut, Robert E., and Paul Resnick. Building Successful Online Communities: Evidence-Based Social Design. MIT Press, 2012.

Marcellino, William, Kate Cox, Katerina Galai, Linda Slapakova, Amber Jaycocks, and Ruth Harris, Human-machine detection of online-based malign information. Santa Monica, CA: RAND Corporation, 2020. https://www.rand.org/pubs/research_reports/RRA519-1.html.

National Academies of Sciences, Engineering, and Medicine. (2019). “A Decadal Survey of the Social and Behavioral Sciences: A Research Agenda for Advancing Intelligence Analysis”. The National Academies Press. https://www.nap.edu/catalog/25335/a-decadal-survey-of-the-social-and-behavioral-sciences-a

Nielsen, Jakob. “Participation Inequality: The 90-9-1 Rule for Social Features.” Nielsen Norman Group (blog), October 2006. https://www.nngroup.com/articles/participation-inequality/.

Nonnecke, Blair, and Jenny Preece. “Lurker Demographics: Counting the Silent.” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 73–80. CHI ’00. New York, NY, USA: Association for Computing Machinery, 2000. https://doi.org/10.1145/332040.332409.

Preece, Jennifer, and Ben Shneiderman. “The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation.” AIS Transactions on Human-Computer Interaction 1, no. 1 (March 2009): 13–32. https://doi.org/10/gfs4km.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever.

"Language Models Are Unsupervised Multitask Learners." San Francisco, CA, USA: OpenAI, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv, July 2023. https://doi.org/10.48550/arXiv.2307.09288.