Extending Persuasion for Misinformation to Graphs

Safwan Hossain shossain@g.harvard.edu
Aryan Gupta aryan.cs.app@gmail.com

The analysis conducted thus far[1] considers interaction between a platform and a single user. In reality, social media interactions are richer – content shared by a user will be observed by another, who can then decide to further disseminate this within their circle. These interactions can most naturally be captured by a graph G where the nodes represent users and edges represent direct connections between users (follower, friends, etc). In general, such a graph may be disjoint, have cycles, and exhibit other rich properties. If we consider, however, the interaction dynamics a single user i's sharing decision entails, information flow is ordered – information flows from the user to their first degree connections, their second degree connections, and so on. Those nodes disconnected from the user in G will never receive any information user i shares. Therefore, the information flow induced by a single user i's decision can be captured by the considered tree induced by performing a breadth-first search rooted at node i. We formalize this below:

Definition 1 (Social Network Graph)**.** *For a population of n users in a social media platform, let $G = (V, E)$ denote an undirected graph where V represents all users and E, the connectivity between them. For any user $i \in V$, let $R_i = (V_i, E_i)$ denote the Directed Acyclic Graph (DAG) obtained by performing a Breadth-First traversal from node i. V_i denotes the set of users that can be reached from user i, and E_i are the set of directed edges where $e = (j_1, j_2)$ implies user j_2 learned of user i's information from user j_1.*

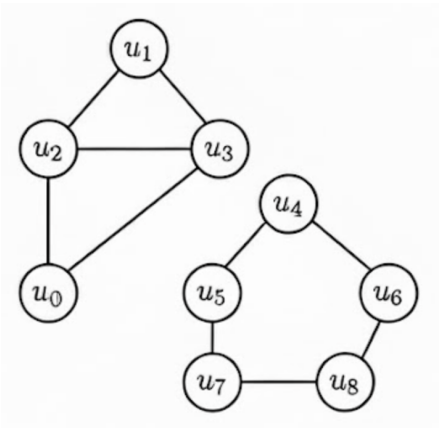

Figure 1: The general social network graph G consisting of 9 users.

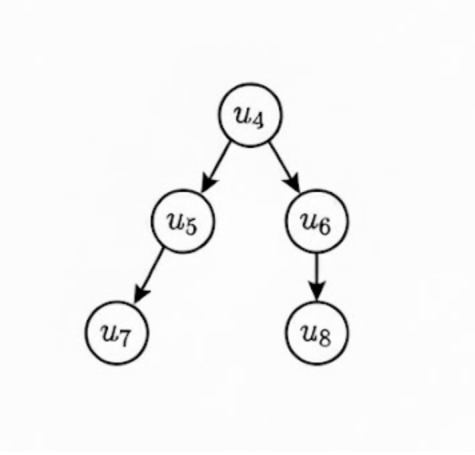

Figure 2: DAG R_4 induced by the connectivity of user 4 in this graph

[1] See [1].

In our original work, we ascribed fixed utilities to the platform for a user sharing or not sharing their content. In reality, this utility endogenously arises from how information shared by a user spreads in the network. The utilities involved in a very well-connected user sharing/not sharing is naturally much more heightened than a poorly connected one. We now look to model this dynamic explicitly. To model the information transmission, we need to precisely define the *state*, *belief*, and intra-user belief transitions. Formally:

Definition 2 (State)**.** *For a user i who created a content c, let ζ denote any objective features about this content – i.e. misinformation/veracity, topic, length, language features, etc. Let κ_i denote any subjective (with respect to the user) features of this content – perceived veracity, peer perception/validation, etc. For the same content, κ_i may not be the same as κ_j. We denote a state $\theta_i = (\zeta, \kappa_i)$.*

Definition 3 (Actions and Utilities)**.** *User i faced with (created or received) content c can choose to either "share" or "not share" – $a_i \in \{0, 1\}$. User and platform utility for the action depends on both action and state. For user, it's given by $w_i(\theta_i, a_i)$ and for platform it is $u(\theta_i, a_i)$* [2]*.*

We assume that for content c created by user i, the platform knows the realized $\theta_i = (\zeta, \kappa_i)$. Further, both user i and the platform share a common prior $\mu_i(\theta)$ over the distribution of states. For all users connected to agent i, the objective features ζ of content c do not change as it propagates; the subjective features, however, change as it is presented to different users. The dynamics of these subjective feature changes is given by the transition kernel P.

Definition 4. *For a given content generated by a user i, consider any two downstream users j_1, j_2 with $e = (j_1, j_2) \in E_i$. Then $P_{j_1,j_2}(\theta_{j_2}|\theta_{j_1})$ represents the conditional probability distribution over user j_2's features given some realization of j_1's features. For brevity, we shall just use $P(\theta_{j_2}|\theta_{j_1})$ hereinafter, and recall that only κ can be different – $\theta_{j_1} = (\zeta^*, \kappa_{j_1}) \implies \theta_{j_2} = (\zeta^*, \cdot)$.*

Lastly, we come to the issue of platform intervention. We generally consider the platform intervening by way of Bayesian Persuasion signaling to the content creator. While we do not consider signaling downstream users who receive content and decide to further propagate or not, our method can be immediately extended to handle this case; we expand on this case as pertinent. We can now formally outline the overall interaction induced by a user i creating content and deciding to disseminate it (or not). The platform can signal not only the content creator, but also downstream users. Note that $\text{pa}(j_2) = j_1$ is used to denote that j_1 is the parent of j_2 and $\text{ch}(j_1) = j_2$ that j_2 is a child of j_1.

1. Platform pre-computes the DAG R_i induced by a user i's network connectivity.
2. User i drafts content c, whose features $\theta_i = (\zeta, \kappa_i)$, the platform observes (perhaps noisily)[3].
3. Platform commits to a signaling scheme $\pi(s|\theta)$.
4. User observes signal s and computes $\mu_i(\theta_i|s)$ and takes optimal action.
5. If user i decides to share:
   (a) Content c and state $\theta_i = \theta^*$ is observed by all first-degree connections.

[2] In practice, the platform utility should only depend on the objective features ζ since we are directly modeling network effects here. Similarly, the user utility would only depend on the subjective features κ.

[3] We do not consider the noisy part in this setup – as we will show, the results from noisy observation are orthogonal and can directly fit our results here.

(b) For an agent j with $\text{pa}(j) = i$, their belief for this content is now $\mu_j(\theta_j) = P(\theta_j|\theta_i = \theta^*)$. They take optimal action according to this belief[4].

(c) The platform receives utility due to the action of this downstream agent.

(d) For any first-degree connections who shared, their connections (user i's 2^{nd} degree connections) would follow a similar procedure recursively.

Theorem 1. *The platform's optimal signaling policy can be efficiently computed in polynomial time.*

Proof. Before resolving the optimal signaling for user i, we ask a simpler question. For any signaling scheme $\pi(s|\theta_i)$ the platform chooses for user i, how can we compute the expected utility under this propagation model? Given the recursive nature of the interaction, answering this requires a few intermediate definitions. First, for any downstream user j, we define $a_j^*(\theta_{\text{pa}(j)})$ as the optimal action taken by user j when they observed content c and parent feature $\theta_{\text{pa}(j)}$. For agent i, it is simply the optimal action given the posterior belief induced by signal s of signaling scheme π. Mathematically:

$$a_i^*(s) = \arg\max_{a_i \in \{0,1\}} \sum_{\theta_i} \mu_i(\theta_i)\pi(s|\theta_i)w_i(a_i, \theta_i) \tag{1}$$

$$a_j^*(\theta_{\text{pa}(j)}) = \arg\max_{a_j \in \{0,1\}} \sum_{\theta_j} P(\theta_j|\theta_{\text{pa}(j)})w_j(a_j, \theta_j) \tag{2}$$

Next, we define our core recursive function. Let $f_j(\theta_{\text{pa}(j)}, \lambda)$ denote the platform expected utility from the subtree rooted at user j, given utility maximizing decisions from all participants. Formally:

$$f_j(\theta_{\text{pa}(j)}, \lambda) = \begin{cases} \sum_{\theta_j} u(\theta_j, a_j^*(\theta_{\text{pa}(j)})) & j = \text{leaf node} \\ \sum_{\theta_j} \left[u(\theta_j, a_j^*(\theta_{\text{pa}(j)})) + \lambda \mathbb{1}[a_j^*(\theta_{\text{pa}(j)}) = 1] \sum_{\ell \in \text{ch}(j)} f_\ell(\theta_j, \lambda) \right] & \text{j} \neq \text{leaf node} \end{cases} \tag{3}$$

With this, we now define the platform utility due to using a signaling scheme π_i for user i's content as the expected cumulative discounted utility due to their decision: $u_i(\pi_i; \mu_i, \lambda)$.

$$u(\pi_i; \mu_i, \lambda) = \sum_{\theta_i} \sum_{s} \mu_i(\theta_i)\pi(s|\theta_i) \left[u(a_i^*(s), \theta_i) + \lambda \mathbb{1}[a_i^*(s) = 1] \sum_{\ell \in \text{ch}(i)} f_\ell(\theta_i, \lambda) \right] \tag{4}$$

The platform goal, thus, is to select a signaling scheme π^* to maximize this expected utility: $\pi^* = \arg\max_\pi u_i(\pi_i; \mu_i, \lambda)$. Despite its unwieldy form, we now show that this optimization problem can be solved in polynomial time. The key to this is observing that the value/continuation function f_j exhibits a memoization property: if one knows the f_ℓ values of all the children of a given node j, then one can directly compute the value of f_j. We thus show in Algorithm 1 how by traversing the DAG in reverse level order, all f_j can be efficiently computed using dynamic programming. This algorithm runs in $O(|\theta||V_i|)$.

Observe that upon running Algorithm 1, we have access to the f_j values for each child of the root i. Since the user i has binary action space, we can use revelation principle to simplify any signaling scheme to be direct and persuasive – that is, signals are action recommendations, and it

[4]If we consider the platform signaling on observed content (not just created content), they would signal here and the optimal action would be influenced by this.

Algorithm 1: Graph Traversal for User i

Compute the DAG rooted at user i. Let the height of the tree be H.
for *each level* h *in* $[H, H-1, \ldots, 2]$ **do**
  for *each node* j *at level* h **do**
    for *each possible parent type* $\theta_{pa(j)}$ **do**
      Let $a_j^*(\theta_{\text{pa}(j)}) = \arg\max_{a_j \in \{0,1\}} \sum_{\theta_j} P(\theta_j|\theta_{\text{pa}(j)}) w_j(a_j, \theta_j)$
      Compute $f_j(\theta_{\text{pa}(j)})$ as in Equation 3 using $a_j^*(\theta_{\text{pa}(j)})$
    end
  end
end

is always incentive compatible for user to follow the recommendation. Observe that the platform utility can be expressed as:

$$u(\pi_i; \mu_i, \lambda) = \sum_{\theta_i} \mu_i(\theta_i)\pi(1|\theta_i) \left[u(1, \theta_i) + \sum_{\ell \in \text{ch}(i)} f_\ell(\theta_i) \right] + \mu_i(\theta_i)\pi(0|\theta_i)u(0, \theta_i) \tag{5}$$

Therefore, we can define a surrogate platform utility u_i' as it pertains to user i: $u_i'(1, \theta) = u(1, \theta_i) + \sum_{\ell \in \text{ch}(i)} f_\ell(\theta_i)$ and $u_i'(0, \theta) = u(0, \theta)$. With this, the sender optimal persuasion problem devolves to the classic persuasion problem:

$$\text{maximize} \sum_{\theta_i} \sum_{a_i} \mu_i(\theta)\pi(a_i|\theta_i)u_i'(a_i, \theta) \tag{6}$$

$$\text{subject to } \forall (a_i, a_i') : \sum_{\theta} \mu_i(\theta_i)\pi(a_i|\theta_i)[w(a_i, \theta_i) - w(a_i', \theta_i)] \geq 0 \tag{7}$$

Corollary 1 (Informal)**.** *All our existing results on (1) imperfect platform knowledge and (2) performative dynamic carry over to the graphical model under the surrogate platform utilities* $u_i'(a_i, \theta_i)$*.*

□

References

[1] Safwan Hossain, Andjela Mladenovic, Yiling Chen, and Gauthier Gidel. A persuasive approach to combating misinformation, 2024.