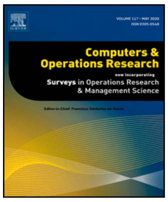
Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor



Stochastic bilevel interdiction for fake news control in online social networks

Kati Moug [a], Siqian Shen [b,*]

[a] School of Industrial and Systems Engineering, Georgia Institute of Technology, United States of America
[b] Department of Industrial and Operations Engineering, University of Michigan, United States of America

ARTICLE INFO

Keywords:
Stochastic network interdiction
Maximum flow
Influence minimization

ABSTRACT

Social media platforms attempt to mitigate and control fake news, using interventions such as flagging posts or adjusting newsfeed algorithms, to protect vulnerable individuals. In this paper, we consider performing intervention actions on specific source nodes or user–user edges in social networks, under uncertain effectiveness of different intervention strategies. We model misinformation from malicious users to vulnerable communities using stochastic network interdiction formulations. Specifically, we minimize the expected number of reachable vulnerable users via stochastic maximum flow, and develop an alternative formulation for handling large-scale social networks based on their topological structures. We derive theoretical results for path-based networks and develop an approximate algorithm for single-edge removal on paths. We test instances of a social network with 23,505 nodes, based on the IMDb actors dataset, to demonstrate the scalability of the approach and its effectiveness. Via numerical studies, we find that characteristics of removed edges change when intervention effectiveness is stochastic. Our results suggest that intervention should target on (i) a smaller set of centrally located edges with nodes that represent communities where regulatory actions are more effective, and (ii) dispersed edges with nodes where intervention has a high chance of failure.

1. Introduction

Peer-to-peer networks have become a major source of news in the contemporary age (see Manjoo, 2017). “Fake news”, purposefully fabricated stories meant to provoke readers, can spread quickly from user to user and platforms have developed strategies for their detection and mitigation (Allcott and Gentzkow, 2017). When misinformation is detected on Facebook, for example, intervention actions include decreasing the post’s ranking in the newsfeed, to lower users’ chance of seeing it, or attaching a “related article” that disputes the post (Iosifidis and Nicoli, 2020). Mitigation strategies like the above occur at the user–user links (or edges) of a social network and can affect a user’s likelihood of either seeing fake news from an account they follow when that person shares. A social media platform may also intervene by blocking nodes in a social network, temporarily stopping misinformation at the source by suspending malicious accounts. Given the vast size of online social networks and the speed at which fake news spreads, it is crucial and challenging to quickly decide the “removal” of source nodes and/or user–user edges, to best control fake news spread to vulnerable users, while considering the stochastic nature of intervention effectiveness.

To our best knowledge, there are two main threads of literature on fake news mitigation: *truth campaigning* and *influence minimization* (Saxena et al., 2022). In truth campaigning, social media networks aim to mitigate the influence of fake news by encouraging the spread of true news, countering misinformation (see, e.g., Budak et al., 2011; He et al., 2012; Farajtabar et al., 2017). On the other hand, influence minimization involves blocking a limited number of edges (Kimura et al., 2009; Kuhlman et al., 2013; Tong et al., 2012) or nodes (Pham et al., 2018; Yao et al., 2015) in a network to minimize the spread of fake news – or other contagions – under a diffusion model, such as the linear threshold or independent cascade model (Kempe et al., 2003). Because we consider direct interventions to stop the spread, the research question in this paper falls into the latter category.

Prior work using influence minimization generally assumes that the effects of regulatory actions are deterministic and fully effective. However, interventions, such as flagging news as “disputed”, may not have the intended effects but sometimes the opposite in practical situations (see, e.g., Saxena et al., 2022). In this paper, we propose a network-interdiction formulation to interrupt the flow of misinformation through a network that has uncertainty in the intervention effectiveness.

Network interdiction is considered as a Stackelberg game (Von Stackelberg and Peacock, 1952) that involves a leader and a follower, where the leader is a player who pays a cost to alter structures of a network, e.g., reducing its arc capacities or blocking its nodes, and the follower is another player who acts after the leader and optimizes their

[*] Corresponding author.
E-mail addresses: kaitlyn.moug@isye.gatech.edu (K. Moug), siqian@umich.edu (S. Shen).

https://doi.org/10.1016/j.cor.2024.106872
Received 5 May 2024; Received in revised form 10 September 2024; Accepted 11 October 2024
Available online 18 October 2024

decisions on the resulting network. The leader aims to alter the network in such a way that the follower's best performance is compromised and thus they "interdict" edges or nodes in the network. Network interdiction examples include maximizing the shortest path (see,e.g., Israeli and Wood, 2002; Song and Shen, 2016) and minimizing the maximum flow of a smuggler who travels from a source node to a destination node, by placing checkpoints on certain arcs (see, e.g., Lei et al., 2018). We refer interested readers to Smith and Song (2020) and Shen (2011) for comprehensive reviews of methods, algorithms, and applications of various network interdiction problems.

In the classical network interdiction setting, the leader and follower both have full knowledge about the problem data and inputs – including the structure of the network and the outcomes of interdiction. As we consider the flow of fake news through a social media network, however, there exist several forms of uncertainty. For instance, if a user sees a piece of information from a malicious source, they may or may not believe it or share it; if the social media network intervenes and attaches a warning to a post, the action may or may not be able to prevent the user from sharing it. With a sufficiently large set of historical data about user behavior, user actions can be modeled by probability distributions. This characteristic of the fake news mitigation problem makes it amenable to the use of stochastic network interdiction for modeling (Cormican et al., 1998).

In this paper, we consider a social media network with vulnerable, malicious, and general users. Our goal is to minimize the expected number of vulnerable users who receive information from malicious users, by removing malicious accounts or intervening on certain user–user edges. (Example actions in practice include, e.g., adapting newsfeed algorithms or adding flags to posts.) We assume that the interventions have a certain user-specific probability of preventing an individual from sharing the information. The distinction of vulnerable users from general users allows us to model the spread of targeted negative information, such as hate speech or content inappropriate for minors. The problem also generalizes to mitigate the flow of fake news to all users in a network if all non-malicious nodes are labeled vulnerable target nodes.

To model this problem, we define and construct a specific supergraph from a given social media network. (We describe the procedures for constructing such a supergraph and provide an illustrative example in Fig. 1 in detail later.) The graph structure allows us to consider both node and edge removal interventions, and also to model the number of vulnerable users reached using a mathematical program. We apply stochastic network interdiction methods (see, e.g., Janjarassuk and Linderoth, 2008) to this graph to minimize the expected number of vulnerable users reached. We also introduce an adaptation of the model for networks with community structures (Girvan and Newman, 2002) to enhance the scalability of our solution approaches.

1.1. Literature review

Wang et al. (2018) consider influence minimization in the context of vulnerable populations, aiming to protect specific target users from fake news with edge blocking. They formulate the problem as an instance of minimum cut-maximum flow when the number of interdicted edges is unrestricted, and utilize greedy algorithms to approach the problem under budget constraints. They only obtain an optimal solution for the unconstrained budget problem, however, while our models optimize a related problem for given budget and under the uncertainty in intervention effectiveness. He et al. (2011) consider both node and edge blocking, and minimize a multi-criteria objective of infection cost and immunization cost. They assume that diffusion is deterministic and takes place in $1 \leq d \leq \infty$ hops, and then provide approximate algorithms for finite and infinite cases. Similarly, their work does not consider stochasticity in information diffusion or intervention effectiveness.

A few prior studies also use a Stackelberg game in constructing models for fake news mitigation. Tanınmış et al. (2020) consider a problem where the leader blocks a set of nodes to minimize influence spread, and then the follower activates nodes to maximize it, following the linear threshold model. They use Sample Average Approximation (SAA) and live-arc representation of the linear threshold model to estimate the solution to the follower's problem, and a greedy heuristic for the leader's problem. Tanınmış et al. (2022) later generalize the problem and develop an improved x-space algorithm for solving min–max bilevel programs that can be used for minimizing misinformation spread. Hemmati et al. (2014) examine a game in which the leader blocks nodes and follower chooses nodes to activate, but follow a deterministic threshold model. They formulate the problem as a two-stage integer program and develop a cutting plane algorithm to solve it. Different from these studies, our paper assumes that the malicious source nodes are known, includes interventions at both the source nodes and user–user edges, and also incorporates uncertain intervention outcomes.

Two-stage stochastic programs are widely used for finding optimal solutions to the influence spread problem. Wu and Küçükyavuz (2018) consider the problem of activating certain nodes in social networks to maximize the number influenced, following independent cascade and linear threshold models, and show that it is a special case of a more general class of two-stage stochastic submodular optimization problems. They develop a delayed constraint generation algorithm for solving the problems optimally when the number of samples is finite. Güney (2019) examines the budgeted influence maximization problem, in which each node has a unique cost to be activated, and develops an integer programming model based on a live-arc representation of the independent cascade model. Güney et al. (2021) later improve the computational efficiency of large-scale influence maximization via maximal covering location design. Song and Dinh (2014) consider the problem of targeting certain edges to minimize misinformation spread given live-arc representations of generic information diffusion cascades. They formulate the problem as a mixed-integer program and solve it with a branch-and-bound algorithm. However, their work does not consider uncertainty in intervention effectiveness.

We also adapt our models for scalability for networks with community structures. A few works consider problems where rumors originate in a particular community. Fan et al. (2013) study the Least Cost Rumor Blocking problem, to protect nodes in communities that neighbor the rumor community. They use a set cover based greedy algorithm in their solution approach. Zheng and Pan (2018) study an extension of the problem, both containing a rumor to a particular community in the graph where it first begins, and constraining its spread within the community itself. They develop a minimum vertex cover based greedy algorithm for the problem. Each of these works mainly focuses on containing the spread within a community and its neighbors, while our adaptation assumes that misinformation spreads quickly through the community. Our interventions instead target spread from a community to another involved in a social network.

1.2. Contributions of the paper

To the best of our knowledge, our work is among the first to utilize stochastic network interdiction models for fake news mitigation. This first allows us to incorporate potential uncertain effects of network intervention actions, as opposed to other studies that assume deterministic interdiction. Second, we adapt the model for scalability, taking advantage of community structure that exists in many social networks. Third, we develop theoretical results about the nature of fake news mitigation with stochastic intervention on single-source paths and use these results to develop an approximation algorithm for our problem on these networks. Finally, we illustrate the use of our model and approaches on a real-world social network data.

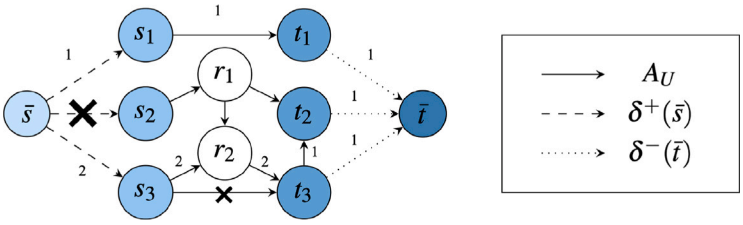


Fig. 1. The graph $G = (N, A)$ built from a social network G_U. The number along side each edge (i, j) is the amount flows on the edge, representing information passed from user i to user j, and edges with no labels have zero flows. Here, system-initiated interdiction is indicated by a large X, and user-initiated interdiction is indicated by a small x.

1.3. Structure of the paper

The remainder of the paper is organized as follows. In Section 2, we discuss detailed notation and problem setup, and then present the mathematical interdiction models for the stochastic fake news control problem. In Section 3, we link our interdiction models to the Independent Cascade Model used in the influence maximization studies. In Section 4, we provide an alternative formulation using a community-based transformation of a social media network, for improving the scalability of our approaches. In Section 5, we focus on path networks and derive insights about solution features and an approximate algorithm for specially structured networks. In Section 6, we conduct extensive numerical studies and present the results. Lastly, in Section 7, we conclude the paper and discuss future research directions.

2. Mathematical models

2.1. Parameters and problem setup

Consider a social network G_U with a set U of users, divided into a set S of malicious source nodes where $S = \{s_i : i = 1, \dots, |S|\}$, a set T of target nodes where $T = \{t_i : i = 1, \dots, |T|\}$, and a set R of intermediary nodes where $R = \{r_i : i = 1, \dots, |R|\}$. The set S represents a group of individuals, each of whom independently shares a piece of fake news with their followers, while T represents the set of their target nodes. This can be a specific group, such as particular targets of hate speech or minor children. The set $R = U \backslash (S \cup T)$, which may be empty, includes users who are neither targets nor sources of misinformation, but who may inadvertently spread misinformation through the network. User nodes are connected in the graph by edges in set A_U. If user j follows user i, then edge $e = (i, j) \in A_U$. Together, these nodes and edges constitute the graph $G_U = (U, A_U)$.

The objective is to minimize the expected number of target nodes that receive the fake news, by interdicting edges or nodes subject to respective budgets. To do so, we formulate the number of target nodes reached as a maximum flow problem on a reconstructed network $G = (N, A)$ built based on the social network G_U. Here, $N = U \cup \{\bar{s}, \bar{t}\}$ is the set of nodes, consisting of a dummy "origin" node $\bar{s}$, a dummy "destination" node $\bar{t}$ and all the original nodes in U, and the edge set is $A = A_U \cup \delta^+(\bar{s}) \cup \delta^-(\bar{t})$, where $\delta^+(\bar{s}) = \{(\bar{s}, s) : s \in S\}$ is the set of edges pointing from $\bar{s}$ to each source node in S and $\delta^-(\bar{t}) = \{(t, \bar{t}) : t \in T\}$ is the set of edges pointing from each target node in T to $\bar{t}$. An example can be seen in Fig. 1. In our maximum flow model, arcs in $\delta^-(\bar{t})$ each have capacity 1, while all other arcs are not capacitated. When we maximize the total amount of flows through G, the bottlenecks are the in-arcs of the destination node $\bar{t}$. Due to flow balance constraints, a positive flow on arc $(t, \bar{t}) \in \delta^-(\bar{t})$ implies that fake news eventually reaches target node user $t \in T$. By constraining flow variables to integer values, we have that a particular target user node is reached if and only if the flow from that node to $\bar{t}$ is 1. Note that the predecessors of $\bar{t}$ in G are precisely T, by definition. Thus, optimizing the maximum flow problem yields the number of target nodes reached by fake news.

We assume that fake news flows freely through the network unless interdiction occurs at particular arcs, where the interdiction of arc (i, j) indicates user j not seeing or believing the information shared by user i. In this paper, we differentiate two types of interdictions as the *system-initiated interdiction* versus *user-initiated interdiction*, such that the former refers to actions that the leader takes to stop the spread of fake news, and the latter refers to stopping of the fake news due to users not believing the information but not because of the leader's interdiction. Later, we model the former as decision variables by the leader and model the latter as input parameters indicating the likelihood of users believing fake news from those they follow. Moreover, there are two types of system-initiated interdiction actions that the leader can take to interrupt the flow of fake news. First, the network can remove a certain number of malicious users; by assumption, this is a deterministic action that completely removes targeted users. Second, the network can attempt to "remove" a certain number of user–user edges. The removal of $e = (i, j) \in A_U$ represents an intervention that prevents j from either seeing the news from i, believing the news from i, or wanting to share the news from i. The success of this intervention on arc $e = (i, j)$ is characterized by a Bernoulli random variable $\tilde{\xi}_e$ with a success probability p_e^{ξ}. For user-initiated interdiction, a user may choose not to believe another user with probability p_e^{α}, which is characterized by a Bernoulli random variable $\tilde{\alpha}_e$, independent of system-initiated interdiction, and is an exogenous parameter only depending on specific edge $e = (i, j)$ and thus the characteristics of users i and j. Note that the random variables involved in our model are $\tilde{\xi} = (\tilde{\xi}_e)^{\mathsf{T}} \in \{0, 1\}^{|A_U|}$ and $\tilde{\alpha} = (\tilde{\alpha}_e)^{\mathsf{T}} \in \{0, 1\}^{|A_U|}$. We can enumerate the support of the joint uncertainty as a finite set Ω of scenarios. Each scenario $\omega \in \Omega$ has an associated realization ξ_ω and α_ω, and probability p_ω.

Next, for each scenario $\omega \in \Omega$ and edge $(i, j) \in A$, we define flow variable $y_{ij\omega} \in \mathbb{Z}_+$. For target node $t \in T$, $y_{t\bar{t}\omega} \geq 1$ if misinformation reaches target node t. For user arcs $(i, j) \in A_U$, $y_{ij\omega} \geq 1$ if i shares the misinformation and j receives it, on a directed path to a target node. Finally, for $s \in S$, $y_{\bar{s}s\omega} \geq 1$ if s shares the fake news with their followers, on a directed path to a target node. Otherwise, $y_{ij\omega} = 0$. Then, for each arc $(i, j) \in A_U \cup \delta^+(\bar{s})$, we define system-initiated interdiction variable $x_{ij} \in \{0, 1\}$. For arcs in A_U, this interdiction represents a user–user link intervention. For user source node $s \in S$, $x_{\bar{s}s} = 1$ if the social media network suspends the user.

An illustrative example. We illustrate an interdiction solution and its impacts on graph G in Fig. 1 as an example. The corresponding social media network G_U consists of source users $S = \{s_1, s_2, s_3\}$, target users $T = \{t_1, t_2, t_3\}$, and general users $R = \{r_1, r_2\}$, and social network relationship arcs A_U are drawn with solid lines. The supergraph arcs $\delta^+(\bar{s})$ and $\delta^-(\bar{t})$ are drawn with dashed and dotted lines, respectively. System-initiated interdiction, illustrated by a large X, occurs on arc $(\bar{s}, s_2) \in \delta^+(\bar{s})$, representing removal of malicious source node s_2. In the scenario ω depicted, user-initiated interdiction occurs on $(s_3, t_3) \in A_U$, shown with a small x, i.e., $\alpha_{s_3 t_3 \omega} = 1$. This represents t_3 not believing fake news from malicious source s_3. Note that user-initiated interdiction does not occur on (r_2, t_3), which means t_3 remains willing to believe fake news from general user r_2. For this reason, news flows from s_3 to r_2 to t_3 to t_1, as well as from s_1 to t_1.

2.2. Stochastic network interdiction formulation

For each scenario $\omega \in \Omega$, we denote $f(x, \xi_\omega, \alpha_\omega)$ as the number of nodes who receive fake news, given system-initiated interdiction decision $x \in \{0,1\}^{|A_U \cup \delta^+(\bar{s})|}$ and realizations ξ_ω, $\alpha_\omega \in \{0,1\}^{|A_U|}$. We formulate the stochastic network interdiction model for mitigating fake news as follows.

$$\min_x \sum_{\omega \in \Omega} p_\omega f(x, \xi_\omega, \alpha_\omega) \tag{1a}$$

$$\text{s.t.} \sum_{(i,j) \in A_U} x_{ij} \le B_U, \tag{1b}$$

$$\sum_{(i,j) \in \delta^+(\bar{s})} x_{ij} \le B_S, \tag{1c}$$

$$x_{ij} \in \{0,1\}, \ (i,j) \in A \setminus \delta^-(\bar{t}), \tag{1d}$$

where constraints (1b) and (1c) impose budgets B_U and B_S, on the number of edges in A_U and source nodes in $\delta^+(\bar{s})$ that one can interdict, respectively, with the overall objective being minimizing the expected number of target nodes reached. The inner problem, to solve for $f(x, \xi_\omega, \alpha_\omega)$ in each scenario ω, is presented as follows.

$$f(x, \xi_\omega, \alpha_\omega) = \max_{y,v} v_\omega \tag{2a}$$

$$\text{s.t.} \sum_{j \in \delta^-(i)} y_{ji\omega} - \sum_{j \in \delta^+(i)} y_{ij\omega} = \begin{cases} -v_\omega, & i = \bar{s} \\ 0, & i \in U \\ v_\omega, & i = \bar{t}, \end{cases} \tag{2b}$$

$$y_{ij\omega} \le \begin{cases} |T|(1 - x_{ij}), & (i,j) \in \delta^+(\bar{s}) \\ |T|(1 - x_{ij}\xi_{ij\omega})(1 - \alpha_{ij\omega}), & (i,j) \in A_U \\ 1, & (i,j) \in \delta^-(\bar{t}), \end{cases} \tag{2c}$$

$$y_{ij\omega} \in \mathbb{Z}^+, \ (i,j) \in A, \tag{2d}$$

$$v_\omega \in \mathbb{Z}^+. \tag{2e}$$

Constraints (2b) ensure flow balance at each node in G. Constraints (2c) ensure zero flow on any successfully system- or user-interdicted arcs. These constraints also ensure that the flow from each target node to the sink node is 0 or 1. Note that these bottlenecks ensure that flow is no more than $|T|$ on any particular arc. Thus, our infinite capacity arcs technically have an upper bound $|T|$ on flow.

If the target to sink node flow $y_{t\bar{t}\omega}$ is 1 for a particular optimal solution to (2), the flow balance constraints imply that the target node t is reached in that solution. If the target to sink node flow $y_{t\bar{t}\omega}$ is 0 and t can be reached, then there is a path from the source node to t with infinite capacity arcs (and flow at most $|T| - 1$), and we can add 1 to that path flow, increasing the objective value, a contradiction. Thus, we have that the objective (2a) yields the number of reachable target nodes.

Note that we can omit the integrality constraints in Model (2), and instead only require $y_{ij\omega} \ge 0$ for $(i,j) \in A$ and $v_\omega \ge 0$ if all arc capacities are integer (see Ahuja et al., 1993). For each scenario $\omega \in \Omega$, we define dual variables $h_\omega \in \mathbf{R}^{|A|}$ and $w_\omega \in \mathbf{R}^{|U|}$. We present the equivalent dual reformulation of (2) as below, for which we define $\xi_{\bar{s}s} = 1$ and $\alpha_{\bar{s}s\omega} = 0$ for all $s \in S$ to unify notation for arcs A_U and $\delta^+(\bar{s})$ in constraints (2c) before taking the dual.

$$f(x, \xi_\omega, \alpha_\omega) = \min_{w,h} \sum_{(i,j) \in \delta^-(\bar{t})} h_{ij\omega} + \sum_{(i,j) \in A_U \cup \delta^+(\bar{s})} |T|(1 - x_{ij}\xi_{ij\omega}) \times (1 - \alpha_{ij\omega})h_{ij\omega} \tag{3a}$$

$$\text{s.t.} \ w_{\bar{s}\omega} - w_{\bar{t}\omega} \ge 1, \tag{3b}$$

$$h_{ij\omega} - w_{i\omega} + w_{j\omega} \ge 0, \ (i,j) \in A, \tag{3c}$$

$$h_{ij\omega} \ge 0, \ (i,j) \in A. \tag{3d}$$

Putting together models (1) and (3), we have an outer and inner minimization problem, that allows us to find the system-initiated interdiction decisions that yield the lowest expected number of vulnerable target users that receive fake news.

$$\min_x \sum_\omega p_\omega \left(\sum_{(i,j) \in \delta^-(\bar{t})} h_{ij\omega} + \sum_{(i,j) \in A_U \cup \delta^+(\bar{s})} |T|(1 - x_{ij}\xi_{ij\omega})(1 - \alpha_{ij\omega})h_{ij\omega} \right) \tag{4}$$

s.t. (1b)–(1d),

(3b)–(3d), $\forall \omega \in \Omega$.

Notice the second summation in the objective includes terms of the form $x_{ij}h_{ij\omega}$, the product of two variables. Letting $z_{ij\omega} = x_{ij}h_{ij\omega}$, after applying McCormick inequalities (McCormick, 1976), one can easily show that model (4) is equivalent to the following linear program, which we can directly solve to obtain optimal system-interdiction decisions.

$$\min_x \sum_\omega p_\omega \left(\sum_{(i,j) \in \delta^-(\bar{t})} h_{ij\omega} + \sum_{(i,j) \in A_U \cup \delta^+(\bar{s})} |T|(h_{ij\omega} - z_{ij}\xi_{ij\omega})(1 - \alpha_{ij\omega}) \right) \tag{5a}$$

s.t. (1b)–(1d),

(3b)–(3d), $\omega \in \Omega$,

$$z_{ij\omega} \le x_{ij}, \ (i,j) \in A \setminus \delta^-(\bar{t}), \ \omega \in \Omega, \tag{5b}$$

$$z_{ij\omega} \le h_{ij\omega}, \ (i,j) \in A \setminus \delta^-(\bar{t}), \ \omega \in \Omega, \tag{5c}$$

$$z_{ij\omega} \ge 0, \ (i,j) \in A \setminus \delta^-(\bar{t}), \ \omega \in \Omega. \tag{5d}$$

3. Connections to independent cascade model

In the stochastic network interdiction-based formulation, we assume that misinformation flows through the network freely unless interrupted by user-initiated or system-initiated interdiction. Our goal is to choose the system-initiated interventions to minimize the expected number of vulnerable users reached under these diffusion assumptions. Note that our models can be viewed and related to the information diffusion model known as the independent cascade model.

3.1. Independent cascade model

Within the independent cascade model, a set of nodes are initially "activated", or infected with misinformation. Then, each infected node attempts to infect a neighbor, and is successful with an arc-specific probability. If they are unsuccessful, they do not attempt to infect that neighbor again. This process repeats with the newly infected nodes attempting to infect their neighbors. The propagation continues until no new nodes are infected.

Formally, suppose we have social media network $G_U = (U, A_U)$ described in Section 2.1. Let each edge $(i,j) \in A_U$ have probability parameter p_{ij}^{flow}. Let W_τ be the set of nodes infected with misinformation at time step $\tau = 0, \dots, |U|$. The set of source nodes $S \subset U$ are initially infected with misinformation, i.e., $W_0 = S$. For $\tau = 0, \dots, |U| - 1$, each newly infected node $i \in W_\tau$ attempts to infect uninfected neighbor $j \in \{k : (i,k) \in A_U\} \setminus (\cup_{z \le \tau} W_z)$, and is successful with probability p_{ij}^{flow}. If successful, $j \in W_{\tau+1}$. This iterative process completes when W_τ is empty for some $\tau = 1, \dots, |U|$.

3.2. Source-aware targeted influence minimization with stochastic intervention

Suppose that we have social media network $G_U = (U, A_U)$ described in Section 2.1, and fake news propagates through G_U according to the independent cascade model, where probability of information flow p_{ij}^{flow} is defined to be $1 - p_{ij}^{\alpha}$, for $(i,j) \in A_U$. Suppose that a certain number of source nodes, B_S, can be removed and a certain number of edges, B_U, can be targeted for removal. The success of attempted removal of edge $(i,j) \in A_U$ is a binomial random variable $\tilde{\xi}_{ij}$ with probability p_{ij}^{ξ}. The *Source-Aware Targeted Influence Minimization with Stochastic Intervention* problem is defined as: How we choose a set of source nodes $\bar{S} \subset S$ with $|\bar{S}| \le B_S$ and a set of edges $\bar{A_U} \subset A_U$ with $|\bar{A_U}| \le B_U$ to remove, so that we minimize the expected number of vulnerable users T infected?

Theorem 1. *The optimal solution to Source-Aware Targeted Influence Minimization with Stochastic Intervention is given by (1) and therefore, (5).*

Proof. Models (1) and (5) find the system-initiated interdiction actions that minimize the expected number of vulnerable nodes T reached over the support Ω. We will show that each scenario $\omega \in \Omega$ corresponds to a live arc representation of the independent cascade model with an adaptation to include stochastic interventions.

We first fix the system-initiated interdiction decision x and $\omega \in \Omega$ and have realizations α_ω and ξ_ω. The function $f(x, \xi_\omega, \alpha_\omega)$ finds the number of target nodes T reachable from $S \setminus \{s \in S : x_{\bar{s}s} = 1\}$ given source node deletions, successful system edge deletions, and user-initiated edge deletions in scenario ω, within the social media network G_U. We then construct a subgraph $G_U^\omega(x) = (U^\omega(x), A_U^\omega(x))$ that incorporates these source node and edge removals. Arc $(i,j) \in A_U$ is included in $A_U^\omega(x)$ if i is not a deleted source node under x and $(1 - \alpha_{ij\omega})(1 - \xi_{ij\omega} x_{ij}) = 1$ under x and ω.

Graph $G_U^\omega(x)$ corresponds to a live-arc representation of the independent cascade model that includes uncertainty in edge removal success. To see this, we follow the arguments given by Kempe et al. (2003). Imagine that at some step in the independent cascade model information propagation process, that node i attempts to infect node j. This takes place if information flows and the edge is not removed. Whether information flows is a random event that can be seen as a weighted coin flip with probability p_{ij}^{flow}. Whether the edge is removed is a two-part question that includes (i) whether the edge is targeted and (ii) whether the targeting is successful, and thus is a random event that can be seen as a weighted coin flip with probability p_{ij}^{ξ}. Note that due to the law of total expectation, we can flip this coin even if the edge in question is not targeted by x. Whether we flip these coins at the moment i attempts to infect j or before the process occurs, does not affect the values. Thus, these coin flips, on all edges of G_U, can be seen as static. The number of infected vulnerable nodes in a scenario of a particular set of coin flips can be determined by making a live-arc representation of the graph. An arc is included in this representation if its coin flips and the system-initiated interdiction decision indicate information would propagate, and excluded otherwise. If a path exists in the live-arc representation from a source node to a vulnerable node, then that set of dynamic coin flips would lead to the infection of the vulnerable node. Thus, the expected number of vulnerable nodes can be found by enumerating all coin flip scenarios, finding the probability of each and number of reachable vulnerable nodes for each, and calculating the expectation.

Each scenario $\omega \in \Omega$ and associated graph $G_U^\omega(x)$ corresponds to a particular set of coin flips — where the arc $(i,j) \in A_U^\omega(x)$ if and only if the coin flips and system-initiated interdiction for that arc would lead to information propagating. Thus, the objective function $\sum_{\omega \in \Omega} p_\omega f(x, \xi_\omega, \alpha_\omega)$, that finds the expected number of reachable nodes over graphs equivalent to the live-arc representations, finds the expected number of infected vulnerable nodes under the information cascade model, and Models (1) (5), find the optimal solution. This completes the proof. □

4. Adaptation for scalability

Implementation of our combined minimization model can be computationally expensive for large social media networks. In this section, we give an alternative formulation of the stochastic network interdiction model for a community-based transformation of a social media network.

Suppose that the set U of user nodes in a large social media network G_U are partitioned into k communities, $C_1, \ldots, C_k$, using a cluster analysis or hierarchical clustering method. We define the community graph of the social network $G_U^c = (U^c, A^c)$ as follows. The set of nodes $U^c = \{1, \ldots, k\}$ each represents a community. For $i \neq j \in U^c$, there is an arc $(i,j) \in A^c$ if and only if an arc exists from a user in community C_i to a user in community C_j in A_U. We label the set of all arcs from users in community C_i to users in community C_j as E_{ij}. Note that E_{ij} is a subset of A_U. We assume that p_e^α and p_e^ξ are the same for all $e \in E_{ij}$, and re-define these probabilities p_{ij}^α and p_{ij}^ξ, respectively.

We assume that once a densely connected community becomes "infected" with misinformation, the misinformation spreads quickly throughout that group, with no user-initiated interdiction. Because of the speed of infection in the community, we assume system-initiated interdiction does not take place within communities either. The emphasis of this scalable model is mitigation of fake news spread between communities.

Within the community graph, each community node is labeled as a source, target, or general node, analogously to the social media network. Source community nodes, labeled S^c, are communities that contain malicious source nodes. Target community nodes, T^c, are communities that contain target nodes and no source nodes. We label the set of target nodes contained in community C_i as T_i, for each community target node $i \in T^c$. General community nodes R^c are communities that contain only general nodes. The community supergraph $G^c = (N^c, A^c)$ is defined analogously to the social network supergraph $G = (N, A)$, with $N^c = U^c \cup \{\bar{s}, \bar{t}\}$ and $A^c = A_U^c \cup \delta^+(\bar{s}) \cup \delta^-(\bar{t})$. The one change we make when we define G^c is the capacity of the in-arcs of the dummy target node. Rather than making each capacity 1, the capacity of arc $(i, \bar{t}) \in A^c$ becomes the number $|T_i|$ of vulnerable target nodes in community C_i. Because other arcs in the supergraph remain infinite capacity, if community C_i that contains target nodes is reachable in the community graph, maximum flow will saturate the capacity of the arc $(i, \bar{t})$. Thus, maximizing flow over this supergraph is equivalent to determining the number of target nodes outside of communities with source nodes that receive fake news. Because we assume misinformation spreads within a community immediately, we can add the number of target nodes in communities with source nodes to this objective to get the total number of target nodes reached.

When the system performs interdictions on nodes or edges in the community supergraph, the corresponding action on the original social media supergraph is much larger. System-initiated interdiction of community source node i implies interdiction of $|C_i|$ nodes in the community. Similarly, system-initiated interdiction of edge (i,j) in the community supergraph corresponds to interdiction of $|E_{ij}|$ edges in the original social media supergraph. We adapt the budget constraints on system-initiation interdiction to reflect this. Overall, the minimization problem is adapted as follows.

$$\min_x \sum_\omega p_\omega \left(\sum_{(i,j) \in \delta^-(\bar{t})} |T_i| h_{ij\omega} + \sum_{(i,j) \in A_U^c \cup \delta^+(\bar{s})} |T| (h_{ij\omega} - z_{ij} \xi_{ij\omega})(1 - \alpha_{ij\omega}) \right) \tag{6a}$$

$$\text{s.t.} \sum_{(i,j) \in A_U^c} |E_{ij}| x_{ij} \leq B_U, \tag{6b}$$

$$\sum_{(i,j) \in \delta^+(\bar{s})} |C_j| x_{ij} \leq B_S, \tag{6c}$$

$$z_{ij\omega} \leq x_{ij}, \ (i,j) \in A^c \setminus \delta^-(\bar{t}), \ \omega \in \Omega, \tag{6d}$$

$$z_{ij\omega} \leq h_{ij\omega}, \ (i,j) \in A^c \setminus \delta^-(\bar{t}), \ \omega \in \Omega, \tag{6e}$$

$$w_{\bar{s}\omega} - w_{\bar{t}\omega} \geq 1, \ \omega \in \Omega, \tag{6f}$$

$$h_{ij\omega} - w_{i\omega} + w_{j\omega} \geq 0, \ (i,j) \in A^c, \ \omega \in \Omega, \tag{6g}$$

$$h_{ij\omega} \geq 0, \ (i,j) \in A^c, \ \omega \in \Omega, \tag{6h}$$

$$z_{ij\omega} \geq 0, \ (i,j) \in A^c \setminus \delta^-(\bar{t}), \ \omega \in \Omega, \tag{6i}$$

$$x_{ij} \in \{0,1\}, \ (i,j) \in A^c \setminus \delta^-(\bar{t}). \tag{6j}$$

As the size of the community graph grows, the number of calculations required to enumerate the stochastic distribution of scenarios grows exponentially. In our computational studies, for instance, the community graph has 189 nodes. While this is much reduced from the

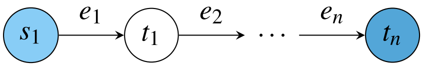


Fig. 2. An illustration of special-structured social media path $\mathcal{P}_n$.

original social network, with 23,505 nodes, enumerating all scenarios is computationally infeasible.

For this reason, we use SAA (Kleywegt et al., 2002) and solve Model (6) M times, each with an independent Monte Carlo sample Ω_m of size n, yielding a candidate solution $\hat{x}_m$. Then, using a large independent Monte Carlo sample $\bar{\Omega}_1$ of size $N > n$, we solve the maximum flow problem for each scenario $\omega \in \bar{\Omega}_1$, estimating the objective value for each candidate solution $\hat{x}_m$. We then choose the candidate solution with lowest objective value, $\bar{x}$, and perform one final test, with a second independent sample, $\bar{\Omega}_2$ of size N, to estimate the mean number of target nodes reached. We obtain statistical upper bound and lower bound of the true optimal objective value through the above procedures.

5. Path networks with stochastic system interdiction

In fake news mitigation, intuition suggests to interdict arcs as close to source nodes as possible — curbing the spread of misinformation before it gains momentum. When does it make sense to delay system-initiated interdiction? To examine this problem more closely, we look at social media networks that can be represented as paths, $\mathcal{P}_n = (U, A_U)$, $n \in \mathbb{N}$, with one malicious source node, s_1, and n target nodes, $\{t_1, \dots, t_n\}$ and user arcs $A_U = \{(s_1, t_1)\} \cup \{(t_i, t_{i+1}) : i = 1, \dots, n-1\}$, as depicted in Fig. 2. Denote each arc $(\cdot, t_i) \in A_U$ as e_i. For conciseness, denote $p^\alpha_{e_i}$ as p^α_i and $p^\xi_{e_i}$ as p^ξ_i. Denote the probabilities of failure to interdict as q^α_i and q^ξ_i, respectively. This network type, while simple, is a subgraph of most real social media networks. Its simplicity helps us gain insight into when delaying system-initiated interdiction makes sense.

First, we assume precisely one arc $e_i \in A_U$ is permitted to be interdicted by the system. When system-initiated interdiction is deterministic, regardless of user-initiated interdiction scenario, earlier interdiction is always better. The first arc in the path whose cost is within budget should be chosen, to minimize the expected number of target nodes reached. If system-initiated interdiction is stochastic, on the other hand, when does it make sense to delay interdiction?

Our integer programming Model (5) provides us one way to determine which arc $e_i \in A_U$ minimizes expected target nodes reached. Another approach involves calculating the expected value, given a particular edge is interdicted by the system. We can repeat this process for all edges in A_U or some subset, if we reduce the size of the search space.

We develop solution methods that utilize this second approach in this section. We start by deriving the explicit form of the expected value for this problem in the following theorem.

Theorem 2. *Suppose that a social media network is described by graph $\mathcal{P}_n$ in Fig. 2, and $j \in \{1, \dots, n\}$. The expected number of target nodes reached when edge e_j is interdicted by the system, denoted $\bar{v}(x_{e_j})$, is given by*

$$\sum_{k=1}^{n} k P(\tilde{v} = k | x_{e_j} = 1), \tag{7}$$

where for $k \in \{1, \dots, n\}$,

$$P(\tilde{v} = k | x_{e_j} = 1) = p_{k+1} \prod_{i=1}^{k} q_i. \tag{8}$$

Here, the probability of failure to interdict arc $i = 1, \dots, n$ is given by

$$q_i = \begin{cases} q^\alpha_i, \ i \neq j, \\ q^\alpha_i q^\xi_i, \ i = j, \end{cases} \tag{9}$$

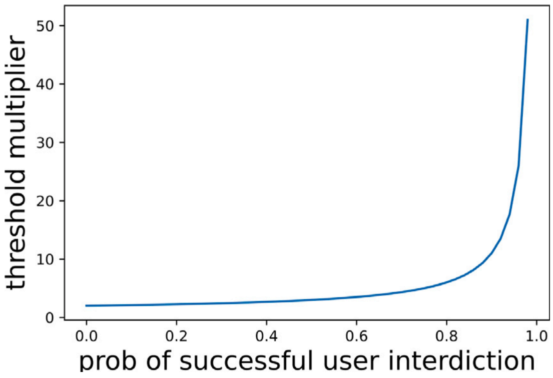


Fig. 3. The threshold multiplier in the delayed system-initiated interdiction threshold, $1 + 1/q^\alpha_2$, as a function of user-initiated interdiction probability p^α_2.

and the probability of successful interdiction of arc $i = 1, \dots, n$ is given by

$$p_i = \begin{cases} p^\alpha_i, \ i \neq j, \\ p^\alpha_i + p^\xi_i - p^\alpha_i p^\xi_i, i = j, \end{cases} \tag{10}$$

while $p_{n+1} = 1$.

Proof. The number of target nodes reached is $k \in \{1, \dots, n\}$ when edges $e_1, \dots, e_k$ are not interdicted by system or user, and edge e_{k+1} is. When $\tilde{v} = n$, interdiction along all arcs has failed. These facts are reflected in Eq. (8) and $p_{n+1} = 1$. For $i \neq j$, the success or failure to interdict comes just from the user, which leads to probabilities p^α_i and q^α_i, respectively. On the other hand, when $i = j$, failure to interdict implies both system and user failed, while success implies either system or user failed. The probabilities in (9)–(10) reflect this, which completes the proof. □

5.1. Two target nodes

Suppose the number of target nodes $n = 2$. In this section, we show that delayed interdiction is optimal when the probability of the second arc's success, p^ξ_2, exceeds a particular threshold, which is the product of the first arc's probability of success, p^ξ_1, and a threshold multiplier, based on the second arc's user-initiated interdiction probability, p^α_2. An illustration of the threshold multiplier, as a function of p^α_2, can be seen in Fig. 3. Notice as the probability of successful user interdiction in the second arc increases, the required size of successful system interdiction probability in the second arc in relation to the first arc grows very quickly. In most cases, early interdiction is preferred over later interdiction.

Theorem 3. *Consider graph $\mathcal{P}_n$ in Fig. 2 with $n = 2$. Suppose the budget allows system-initiated interdiction of either e_1 or e_2. Assume user-initiated interdiction is not deterministic (i.e., $p^\alpha_i < 1$ for $i = 1, 2$). Interdicting arc e_2 minimizes the expected number of target nodes reached if and only if*

$$p^\xi_2 \geq (1 + 1/q^\alpha_2) p^\xi_1. \tag{11}$$

Proof. We prove the result by showing that (11) holds if and only if $\bar{v}(x_{e_1}) - \bar{v}(x_{e_2}) \geq 0$. First, we use Theorem 2 to calculate $\bar{v}(x_{e_1}) - \bar{v}(x_{e_2})$. Note that we have the following probability distributions.

$$P(\tilde{v} = k | x_{e_1} = 1) = \begin{cases} p^\alpha_2 q^\alpha_1 q^\xi_1, \ k = 1, \\ q^\alpha_1 q^\xi_1 q^\alpha_2, \ k = 2. \end{cases} \tag{12}$$

$$P(\tilde{v} = k | x_{e_2} = 1) = \begin{cases} (p^\alpha_2 + p^\xi_2 - p^\alpha_2 p^\xi_2) q^\alpha_1, \ k = 1, \\ q^\alpha_1 q^\alpha_2 q^\xi_2, \ k = 2. \end{cases} \tag{13}$$

Because q_1^α is a positive common factor in all the terms in $\bar{v}(x_{e_1})-\bar{v}(x_{e_2})$, the expression is nonnegative if and only if $\bar{v}(x_{e_1})-\bar{v}(x_{e_2})/q_1^\alpha \geq 0$. Thus, $\bar{v}(x_{e_1}) - \bar{v}(x_{e_2}) \geq 0$ if and only if

$$(p_2^\alpha q_1^\xi - p_2^\alpha + p_2^\xi - p_2^\alpha p_2^\xi) + 2(q_1^\xi q_2^\alpha - q_2^\alpha q_2^\xi) \geq 0 \tag{14}$$

$$\Longleftrightarrow (p_2^\alpha - p_2^\alpha p_1^\xi - p_2^\alpha + p_2^\xi - p_2^\alpha p_2^\xi) + 2q_2^\alpha(p_2^\xi - p_1^\xi) \geq 0 \tag{15}$$

$$\Longleftrightarrow p_2^\alpha(p_2^\xi - p_1^\xi) + p_2^\xi + 2q_2^\alpha(p_2^\xi - p_1^\xi) \geq 0 \tag{16}$$

$$\Longleftrightarrow (1 + q_2^\alpha)(p_2^\xi - p_1^\xi) + p_2^\xi \geq 0 \tag{17}$$

$$\Longleftrightarrow \; - p_1^\xi + q_2^\alpha p_2^\xi - qp_1^\xi \geq 0 \tag{18}$$

$$\Longleftrightarrow p_2^\xi \geq (1 + 1/q_2^\alpha)p_1^\xi. \tag{19}$$

Note that the last line holds because $q_2^\alpha = 1 - p_2^\alpha > 0$. This completes the proof. □

Theorem 3 has a few immediate corollaries.

Corollary 1. *For graph $\mathcal{P}_2$, if $p_1^\xi = p_2^\xi$, then interdiction of arc e_1 is preferred over interdiction of arc e_2, regardless of user-initiated interdiction probabilities.*

Corollary 2. *For graph $\mathcal{P}_2$, the preferred interdiction decision depends on system-initiated interdiction and e_2 user-initiated interdiction probabilities. The decision is independent of the e_1 user-initiated interdiction probability.*

5.2. ϵ-optimal approximation algorithm for path networks

When we consider whether to delay system-initiated interdiction along a path $\mathcal{P}_n$, we compare the expected number of target nodes reached if an edge is interdicted to earlier edges in the path. Independent of system-initiated interdiction, we have user-initiated interdiction. Intuitively, the further we move along the path, delaying system-initiated interdiction, the more likely we are to reach a point where user-initiated interdiction plays the greater role in expected number of target nodes reached. In this section, we show a theoretical proof supports this intuition. The distance between values in the sequence $\{\bar{v}(x_{e_i})\}_{i=1}^n$ decreases as the index increases. We describe this phenomenon in more detail in the theorem below, and follow with an algorithm that uses the result to produce ϵ-optimal solutions to the minimum expected target nodes reached problem for paths of the form $\mathcal{P}_n$.

Theorem 4. *Consider graph $\mathcal{P}_n$ in Fig. 2. For this graph, expected number of nodes reached*

$$|\bar{v}(x_{e_i}) - \bar{v}(x_{e_{i+1}})| \leq \prod_{j=1}^{i-1} q_{e_j}^\alpha \sum_{j=i-1}^{n} j \tag{20}$$

for $i = 2, \ldots, n-1$. The sequence

$$\left\{ \prod_{j=1}^{k} q_{e_j}^\alpha \sum_{j=k}^{n} j \right\}_{k=1}^{n-2} \tag{21}$$

is monotone decreasing.

Proof. First, note that for $k = 1, \ldots, i-2$, $P(\tilde{v} = k|x_{e_i} = 1) = P(\tilde{v} = k|x_{e_{i+1}} = 1) = p_{k+1}^\alpha \prod_{j=1}^k q_j^\alpha$, since no system-initiated interdiction is involved. This implies

$$|\bar{v}(x_{e_i}) - \bar{v}(x_{e_{i+1}})| = \left| \sum_{k=i-1}^{n} k(P(\tilde{v} = k|x_{e_i} = 1) - P(\tilde{v} = k|x_{e_{i+1}} = 1)) \right| \tag{22}$$

$$\leq \sum_{k=i-1}^{n} k \left| P(\tilde{v} = k|x_{e_i} = 1) - P(\tilde{v} = k|x_{e_{i+1}} = 1) \right|. \tag{23}$$

Now, note two cases for the value $P(\tilde{v} = k|x_{e_j} = 1)$. If $j < k+1$, we have that system-initiated interdiction failed, and $P(\tilde{v} = k|x_{e_j} = 1) = p_{k+1} \prod_{l=1}^k q_l^\alpha q_j^\xi$. Otherwise, $P(\tilde{v} = k|x_{e_j} = 1) = p_{k+1} \prod_{l=1}^k q_l^\alpha$. In either case, for $k \geq i-1$, we have

$$\frac{P(\tilde{v} = k|x_{e_j} = 1)}{\prod_{l=1}^{i-1} q_l^\alpha} \in (0, 1), \tag{24}$$

which implies

$$\prod_{j=1}^{i-1} q_{e_j}^\alpha \frac{\left| P(\tilde{v} = k|x_{e_i} = 1) - P(\tilde{v} = k|x_{e_{i+1}} = 1) \right|}{\prod_{j=1}^{i-1} q_{e_j}^\alpha} \leq \prod_{j=1}^{i-1} q_{e_j}^\alpha. \tag{25}$$

This, along with (23), yields the result. □

We utilize Theorem 4 to develop the following ϵ-optimal algorithm.

Algorithm 1 PathInterdict

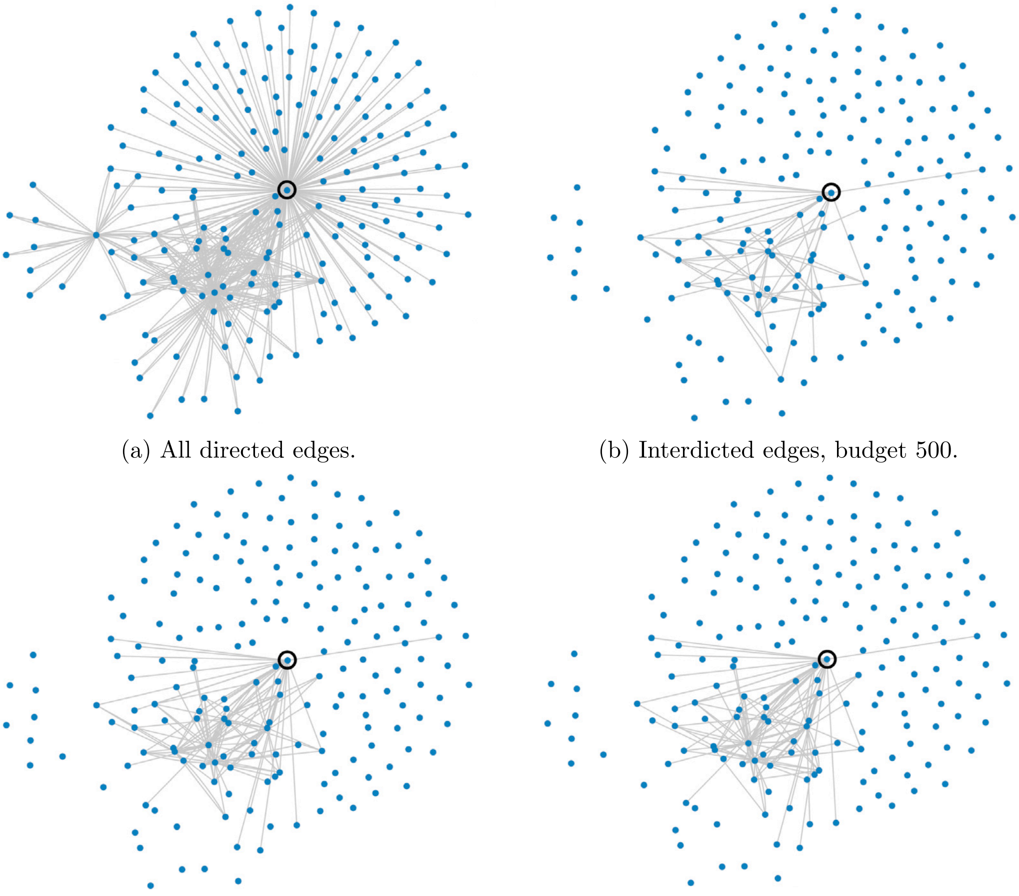
```
Input: Optimal value tolerance $\epsilon$. Path graph $\mathcal{P}_n$.
Output: $\epsilon$-optimal interdiction solution.
Initialize $q := q_{e_1}^\alpha$ and $C := n(n+1)/2$.
for $k = 1, \ldots, n-2$ do
    if $qC < \epsilon/(n-k)$ then
        break
    end if
    Update $q := q * q_{k+1}^\alpha$ and $C := C - k$.
end for
Find expected number of target nodes reached $\bar{v}(x_{e_j})$ for $j = 1, \ldots, k+1$.
return $\text{argmin}_{j=1,\ldots,k+1} \bar{v}(x_{e_j})$.
```

Note that Algorithm 1 checks if the RHS of (20) is less than the tolerance ϵ divided by an upper bound on the remaining number of edges, $n-k$. Because sequence (21) is monotone decreasing, when this line is satisfied for some k, the difference between $\bar{v}(x_{e_{k+1}})$ and $\bar{v}(x_{e_j})$ is less than ϵ for all $j \geq k+2$.

6. Computational results

In Section 4, we develop an adaptation of the stochastic network interdiction model by taking the advantage of the community structure that exists in many social networks and discuss the use of SAA to further reduce computational time. To test the effectiveness of these approaches, we conduct numerical studies in this section by constructing a social network with 23,505 nodes using a subset of the actors born 1990 or later in the IMDb dataset (IMDb, 2023). In the graph, an edge exists between two actors if they have worked on a project (e.g., a movie) together. We use the largest connected component, containing about 90% of original nodes, to construct the social network. To construct the corresponding community graph, we use Clauset-Newman-Moore greedy modularity maximization (Clauset et al., 2004), implemented by NetworkX (Hagberg et al., 2008), partitioning the nodes into 189 communities.

We test different values for the following parameters: number of community source nodes, budget for interdicting edges, budget for interdicting source nodes, and type of random behavior. Specifically, we consider 1, 2, or 3 community source nodes, selecting the largest communities, and test $B_U = 100, 200, 300, \ldots, 1500$ for edge interdiction (with $B_S = +\infty$), and then $B_S = 100, 150, 200, 250, 300$ for source node interdiction (with $B_U = +\infty$). For the user or system behavior types, we assume three general cases, being "Stochastic User" (with $p^\xi = 1$), "Stochastic System" (with $p^\alpha = 0$), and "Both Stochastic" (with both $0 < p^\xi < 1$ and $0 < p^\alpha < 1$), respectively. We summarize the specific choices of p^ξ- and p^α-values in Table 1 for each case. Note that each probability combination will affect the scenarios in the Monte Carlo sampling and the SAA model later. Here, p^α being larger indicates that users in the system have more variability believing fake news, and p^ξ being larger indicates that interdicting edges/source nodes has more uncertainty in its effects.



(a) All directed edges.

(b) Interdicted edges, budget 500.

(c) Interdicted edges, budget 1000.

(d) Interdicted edges, budget 1500.

Fig. 4. Community graph with $|S^c| = 1$ and stochastic users.

Table 1
User and system behavior parameters.

	p^α	p^ξ
Stochastic user	0.1, 0.2, …, 0.5	1
Stochastic system	0	0.5, 0.6 …, 1
Both stochastic	0.1, 0.2, …, 0.5	0.5, 0.6 …, 1

To solve the optimal solution of the community structure model (6), we perform SAA with $M = 5$ random trials of Monte Carlo samples and 100 scenarios in each trial of samples to yield five candidate solutions. We then estimate the objective value of each candidate solution with an independent Monte Carlo sample with 1000 scenarios, solving the maximum flow problem for each scenario and taking the average. We choose the candidate solution with minimum expected number of target nodes reached. Finally, with the chosen candidate solution, we perform these second-phase SAA tests again with a second independent Monte Carlo sample, also with 1000 scenarios, solving the maximum flow problem for each scenario and taking the average, to estimate the objective value of the chosen candidate solution.

All the computational tests were conducted on a computer with an Intel Core E5-2630 v4 CPU 2.20 GHz and 128 GB of RAM. We use Python and Gurobi 8.1 for solving all the optimization models. For clarity and conciseness, we only present results of representative combinations of parameter settings in the main paper.

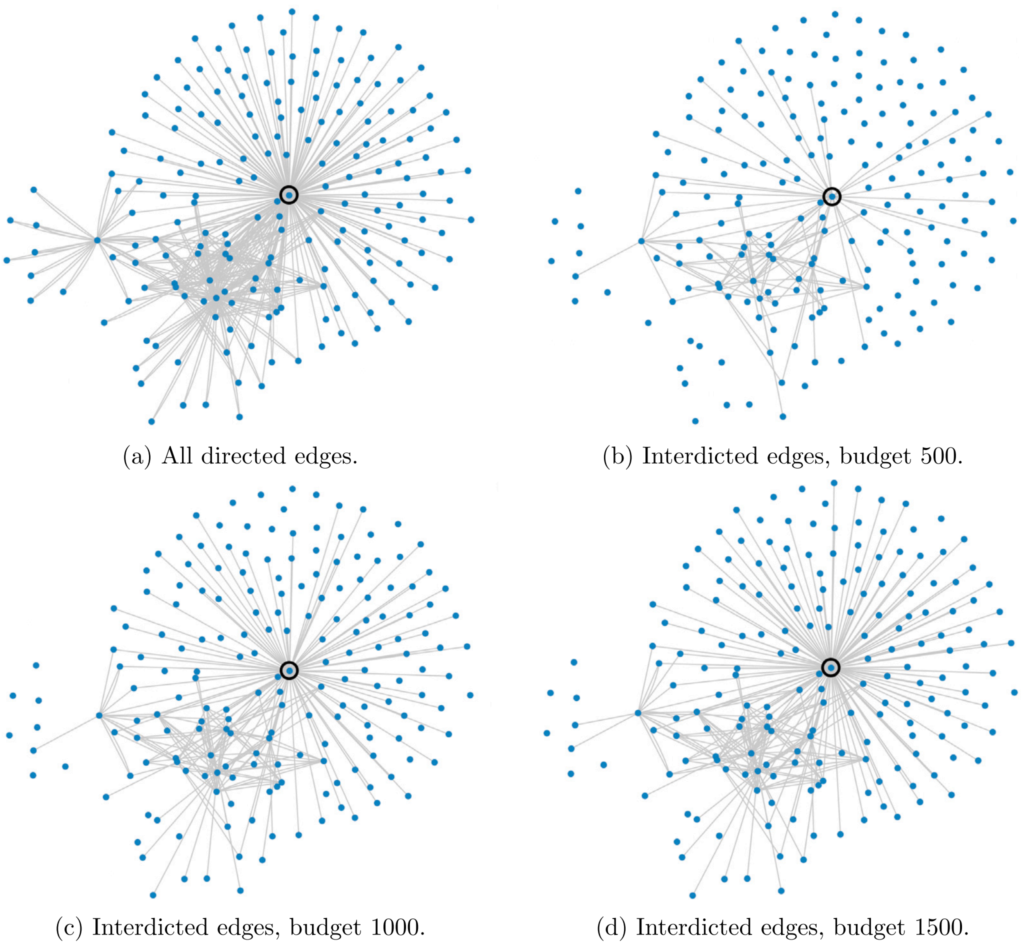
6.1. Performance of the solution methods

For all test instances, we find the SAA optimality gap is between −0.5% and 2.5%. We give the final objective value and various characteristics of interdicted edges for the test instances with $|S^c| = 1$ in Table 2. (We exclude those with $|S^c| = 2$ and 3 for conciseness because the result patterns are similar.)

In Table 2, we report instances with $B_U = 100, 200, 500, 1000, 1500$ and three combinations of stochastic behavior instances ($p^\alpha = 0, p^\xi = 0.5$), ($p^\alpha = 0.5, p^\xi = 0.5$), ($p^\alpha = 0.5, p^\xi = 1$), each representing "Stochastic System", "Both Stochastic", and "Stochastic User", respectively. This is because the small increments of B_U (i.e., from 100 to 200) do not lead to significant result changes (as shown in Table 2), which are also observed for all $B_U = 100, 200, \ldots, 1500$, and thus we only show the representative cases with $B_U = 100, 200, 500, 1000, 1500$. The patterns of solutions are similar for different (p^α, p^ξ) combinations in between the values 0 and 1, and thus we only present results associated with the three general cases described in Table 1 and use 0.5 as the probability value.

From Table 2, the expected number of target nodes reached is the highest when system-initiated interdiction is stochastic and user-initiated interdiction is deterministically zero, and the lowest when the system is deterministically effective and users are stochastically participating in self-interdiction, for fixed budget. Increasing the budget has the greatest effect on the stochastic users case, where system-initiated interdiction is deterministically effective.

Next we investigate what types of edges the mixed-integer program chooses and the properties of these edges. Recall that each node in U^c


(a) All directed edges.

(b) Interdicted edges, budget 500.

(c) Interdicted edges, budget 1000.

(d) Interdicted edges, budget 1500.

Fig. 5. Community graph with $|S^c| = 1$ and both stochastic users and system.

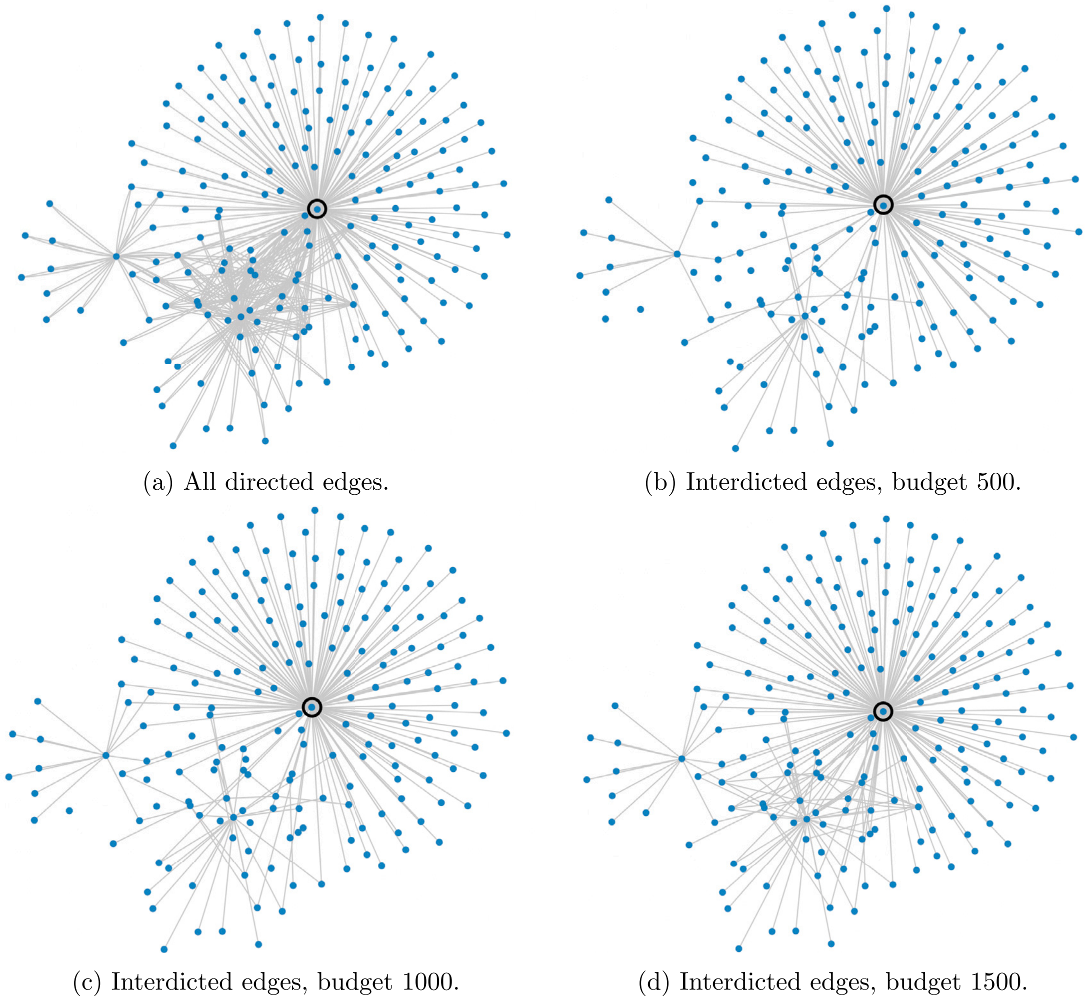
Table 2
Objective values and interdicted edge characteristics for instances with $|S^c| = 1$.

Budget B_U	Stoch. Behavior	Obj. Val.	Interdicted edges		
			Median comm nodes	Mean cost	Total number
100	$p^\alpha = 0,\ p^\xi = 0.5$	15 218.3	3	1.8	132
	$p^\alpha = 0.5,\ p^\xi = 0.5$	13 925.5	293	4.1	58
	$p^\alpha = 0.5,\ p^\xi = 1$	12 149.2	573	6.9	42
200	$p^\alpha = 0,\ p^\xi = 0.5$	15 191.7	3	1.8	138
	$p^\alpha = 0.5,\ p^\xi = 0.5$	13 892.1	271	4.2	62
	$p^\alpha = 0.5,\ p^\xi = 1$	11 167.8	599	7.0	43
500	$p^\alpha = 0,\ p^\xi = 0.5$	14 912.6	6	3.2	154
	$p^\alpha = 0.5,\ p^\xi = 0.5$	13 784.8	229	4.7	106
	$p^\alpha = 0.5,\ p^\xi = 1$	8905.6	690	7.4	68
1000	$p^\alpha = 0,\ p^\xi = 0.5$	14 774.6	6	5.5	182
	$p^\alpha = 0.5,\ p^\xi = 0.5$	13 465.2	140	5	200
	$p^\alpha = 0.5,\ p^\xi = 1$	5737.8	779	7.8	129
1500	$p^\alpha = 0,\ p^\xi = 0.5$	14 717.4	11	6.4	230
	$p^\alpha = 0.5,\ p^\xi = 0.5$	13 346.2	97	6	250
	$p^\alpha = 0.5,\ p^\xi = 1$	3704.8	490	12.4	121

represents a community of nodes in U, whereas each arc (i,j) in A_U^c represents a set of edges $E_{ij} \subset A_U$. In Table 2, in the "Median Comm Nodes" column, we examine the size of community nodes involved in interdicted edges for each test instance. To do so, we take the median of a vector having two components for each interdicted arc, (i,j), with values $|C_i|$ and $|C_j|$, respectively, and exclude any components associated with S^c. Notice that a similar pattern exists for each budget. For all cases with enough interdiction budget (e.g., $B_U = 500, 1000, 1500$), when only the system is stochastic and users deterministically do not self-interdict, the median community size is very low — ranging from 6 to 11. Here, interdiction targets a larger number of smaller communities. On the other hand when only users are stochastic, and system-initiated interdiction is deterministically effective, the median community size is high — ranging from 490 to 779. Here, interdiction tends to target larger communities. When both are stochastic, we see a median between these two. This pattern suggests that centralized, directed action is more effective when users are stochastic, while widespread interdiction has better performance when the system is stochastic.

In Table 2, we also examine the average cost of interdicted edges in the community supergraph. Recall that the cost of edge (i,j) is $|E_{ij}|$, the number of edges from community C_i to community C_j in the original social media network. Notice that for fixed budget, no clear pattern exists distinguishing the stochastic system and both stochastic cases, while stochastic users generally always have a higher mean cost and fewer interdicted edges. This suggests that the stochasticity of system-initiated interdiction has a clearer effect on edge cost than the stochasticity of the user. When the system is deterministically effective, in the stochastic users case, there are fewer interdicted edges with higher cost. When the effectiveness of the system is stochastic, regardless of whether users self-interdict, there tend to be more interdicted edges with lower cost. It depends on how stochastic system and both stochastic compare, however.

Also notice in Table 2 that no clear pattern exists for the impact of increasing budget on number of edges interdicted and mean edge cost. For stochastic system, increasing the budget leads to increases in both, but not linearly. At the same time, for stochastic users with



(a) All directed edges.

(b) Interdicted edges, budget 500.

(c) Interdicted edges, budget 1000.

(d) Interdicted edges, budget 1500.

Fig. 6. Community graph with $|S^c| = 1$ and stochastic system.

$B_U = 100, 200, 500, 1000, 1500$, mean edge cost is 6.9, 7.0, 7.4, 7.8, and 12.4, and mean number of edges is 68, 129, and 121. This suggests a greedy heuristic might not perform as well as the optimization model.

We visualize interdicted edge characteristics in Figs. 4–6. In each figure, part (a) shows the community supergraph G^c with directed edges. We exclude dummy nodes $\bar{s}$ and $\bar{t}$ to simplify the visualization. The community source node is circled; all other nodes are community target nodes. Then, in parts (b)–(d) of each figure, we show the edges that are interdicted by the optimization model, for budgets 500, 1000, and 1500, respectively.

When only users are stochastic, and system-initiated interdiction is deterministically effective, as seen in Fig. 4, the interdicted edges are clustered in a densely connected area. When only the system is stochastic, and users deterministically do not self-interdict, as seen in Fig. 6, the interdicted edges are spread throughout the graph. When both are stochastic, as seen in Fig. 5, we see a pattern between these two extremes, with some interdicted edges in the densely connected area and some spread throughout the graph.

When we interdict source nodes directly by varying $B_S = 100, 150, 200, 250, 300$ instead of interdicting edges, we observe similar results as the above ones, in terms of the interdiction effectiveness as B_S increases, and thus we omit the detailed tables and solution patterns for conciseness. Similarly, when B_S only increases by 50, the effects are negligible and because the number of nodes is significantly smaller than the number of edges, we vary B_S up to 300. In practice, node interdiction often corresponds to temporally or permanently deleting users' accounts, and thus will be more expensive and difficult to implement as compared to edge interdiction.

6.2. Performance on original social network

In this section, we examine performance of the solution derived from community supergraph G^c, when interdiction is applied to the original social network supergraph, G. While the size of G makes the first stage of SAA computationally infeasible for Model (5), we can estimate how our solution for Model (6) performs on the larger graph.

Note that the second Monte Carlo sample with 1000 scenarios used to estimate the expected number of target nodes reached has realizations of α and ξ based on the community supergraph G^c. This indicates that all edges in E_{ij}, with one node in community C_i and another in C_j in the original graph G, have the same realization of user-initiated interdiction and success of system-initiated interdiction for particular scenario ω. If we apply this assumption to G, and use the same Monte Carlo sample, the expected number of target nodes reached in G is equal to the final objective value calculated by our solution methods.

If we generate a new set of samples to estimate Ω for each test instance, where each edge in E_{ij} can have its own realization of user-initiated interdiction and system-initiated interdiction success in a particular scenario, there is no guarantee that the solutions chosen by (6) lead to the same expected number of target nodes reached. Tests with independent Monte Carlo samples of size 20 found differences of up to 37% from the final objective value. This indicates that our community-based model is best used in situations in which user–user edges between communities are affected similarly by system-initiated interdiction and user-initiated self-interdiction.

7. Conclusion

In this paper, we considered how to mitigate the spread of misinformation or hate speech with limited resources available and uncertainty in intervention effect. We developed a model based on stochastic network interdiction that minimizes the expected maximum flow on a supergraph of a social media network. This supergraph is designed to make equivalent maximum flow and the number of vulnerable users reached in the social media network. We derived an integer programming formulation that uses the dual of maximum flow to create a combined minimization problem, as well as a community-based adaptation for scalability. We derived theoretical results about the nature of fake news mitigation in chained social media networks, with one source node and n target nodes. In particular, we discovered that a threshold exists for delaying system-initiated interdiction when $n = 2$, based on the probabilities of system-initiated success and the final arc's probability of user self-interdiction, independent of the first arc's probability of user-initiated success. We also found that as system-initiated interdiction is delayed further and further along a path of length n, the absolute difference in expected number of target nodes reached decreases. This allowed us to develop an ϵ-optimal algorithm for single-arc system interdiction of these networks.

We further applied our community-based mixed integer program (6) to a large social network. We found that when systemic actions are deterministically effective and users stochastically self-interdict, solutions given by our methods tend to target centralized edges that connect nodes representing large communities. On the other hand, when systemic actions have some probability of failure and users deterministically do not self-interdict, solutions tend to include arcs spread throughout the graph that connect nodes representing smaller communities. Interdiction actions chosen using Model (6) yield the same expected number of targeted users reached in the original social network supergraph as the final objective value given by the solution methods when the edges between each pair of communities share user-initiated interdiction and success of system-initiated interdiction realizations. However, such a result does not generalize to the case when each edge has an independent realization of the stochastic parameters in the original social network supergraph. Future research can consider how to develop scalable methods based on community structure that incorporate this characteristic. Our numerical studies show that depending on specific probabilities of system and user behavior uncertainties, the interdiction results and solutions could be very different, and thus the stochastic programming approach requires perfect information and full knowledge of p^{α} and p^{ξ}. In practice, one can only access the stochastic system and user behavior via historical data, which could be limited, and thus the inferred p^{α} and p^{ξ} may not be exact and accurate. As a result, for future research, one can explore alternative models and algorithms to the stochastic programming approach we consider in this paper, e.g., robust or distributionally robust optimization models, where the distribution of uncertain parameter is ambiguously known, and we optimize the outcomes of interdiction decisions against the worst case realization of the uncertain interdiction outcome or probabilities p^{α}, p^{ξ}. For example, Sadana and Delage (2023) study the effectiveness of randomization in interdiction games with an interdictor who is risk and ambiguity averse, and consider a distributionally robust network interdiction game where the interdictor randomizes over the feasible interdiction plans in order to minimize the worst-case conditional value-at-risk of the flow.

CRediT authorship contribution statement

Kati Moug: Writing – original draft, Visualization, Validation, Software, Formal analysis, Data curation. **Siqian Shen:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Data availability

Data will be made available on request.

References

Ahuja, R., Magnanti, T., Orlin, J., 1993. Network Flows: Theory, Algorithms, and Applications. Prentice Hall.

Allcott, H., Gentzkow, M., 2017. Social media and fake news in the 2016 election. J. Econ. Perspect. 31 (2), 211–236.

Budak, C., Agrawal, D., El Abbadi, A., 2011. Limiting the spread of misinformation in social networks. In: Proceedings of the 20th International Conference on World Wide Web. pp. 665–674.

Clauset, A., Newman, M.E., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E 70 (6), 066111.

Cormican, K.J., Morton, D.P., Wood, R.K., 1998. Stochastic network interdiction. Oper. Res. 46 (2), 184–197.

Fan, L., Lu, Z., Wu, W., Thuraisingham, B., Ma, H., Bi, Y., 2013. Least cost rumor blocking in social networks. In: 2013 IEEE 33rd International Conference on Distributed Computing Systems. IEEE, pp. 540–549.

Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., Zha, H., 2017. Fake news mitigation via point process based intervention. In: International Conference on Machine Learning. PMLR, pp. 1097–1106.

Girvan, M., Newman, M.E., 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. 99 (12), 7821–7826.

Güney, E., 2019. On the optimal solution of budgeted influence maximization problem in social networks. Oper. Res. 19 (3), 817–831.

Güney, E., Leitner, M., Ruthmair, M., Sinnl, M., 2021. Large-scale influence maximization via maximal covering location. European J. Oper. Res. 289 (1), 144–164.

Hagberg, A., Swart, P., Chult, D.S., 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

He, J., Liang, H., Yuan, H., 2011. Controlling infection by blocking nodes and links simultaneously. In: Internet and Network Economics: 7th International Workshop, WINE 2011, Singapore, December 11-14, 2011. Proceedings 7. Springer, pp. 206–217.

He, X., Song, G., Chen, W., Jiang, Q., 2012. Influence blocking maximization in social networks under the competitive linear threshold model. In: Proceedings of the 2012 SIAM International Conference on Data Mining. SIAM, pp. 463–474.

Hemmati, M., Smith, J.Cole., Thai, M.T., 2014. A cutting-plane algorithm for solving a weighted influence interdiction problem. Comput. Optim. Appl. 57, 71–104.

IMDb, 2023. Accessed on February 28 2023.

Iosifidis, P., Nicoli, N., 2020. The battle to end fake news: A qualitative content analysis of Facebook announcements on how it combats disinformation. Int. Commun. Gaz. 82 (1), 60–81.

Israeli, E., Wood, R.K., 2002. Shortest-path network interdiction. Networks 40 (2), 97–111.

Janjarassuk, U., Linderoth, J., 2008. Reformulation and sampling to solve a stochastic network interdiction problem. Networks 52 (3), 120–132.

Kempe, D., Kleinberg, J., Tardos, É., 2003. Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 137–146.

Kimura, M., Saito, K., Motoda, H., 2009. Blocking links to minimize contamination spread in a social network. ACM Trans. Knowl. Discov. Data (TKDD) 3 (2), 1–23.

Kleywegt, A.J., Shapiro, A., Homem-de Mello, T., 2002. The sample average approximation method for stochastic discrete optimization. SIAM J. Optim. 12 (2), 479–502.

Kuhlman, C.J., Tuli, G., Swarup, S., Marathe, M.V., Ravi, S., 2013. Blocking simple and complex contagion by edge removal. In: 2013 IEEE 13th International Conference on Data Mining. IEEE, pp. 399–408.

Lei, X., Shen, S., Song, Y., 2018. Stochastic maximum flow interdiction problems under heterogeneous risk preferences. Comput. Oper. Res. 90, 97–109.

Manjoo, F., 2017. Can Facebook fix its own worst bug? N. Y. Times Mag. 24, 2017.

McCormick, G.P., 1976. Computability of global solutions to factorable nonconvex programs: Part I – convex underestimating problems. Math. Program. 10 (1), 147–175.

Pham, C.V., Phu, Q.V., Hoang, H.X., 2018. Targeted misinformation blocking on online social networks. In: Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19-21, 2018 Proceedings, Part I. Springer, pp. 107–116.

Sadana, U., Delage, E., 2023. The value of randomized strategies in distributionally robust risk-averse network interdiction problems. INFORMS J. Comput. 35 (1), 216–232.

Saxena, A., Saxena, P., Reddy, H., 2022. Fake news propagation and mitigation techniques: A survey. Princ. Soc. Netw.: New Horiz. Emerg. Chall. 355–386.

Shen, S., 2011. Reformulation and Cutting-Plane Approaches for Solving Two-Stage Optimization and Network Interdiction Problems (Ph.D. thesis). University of Florida.

Smith, J.C., Song, Y., 2020. A survey of network interdiction models and algorithms. European J. Oper. Res. 283 (3), 797–811.

Song, Y., Dinh, T.N., 2014. Optimal containment of misinformation in social media: A scenario-based approach. In: Combinatorial Optimization and Applications: 8th International Conference, COCOA 2014, Wailea, Maui, HI, USA, December 19-21, 2014, Proceedings 8. Springer, pp. 547–556.

Song, Y., Shen, S., 2016. Risk-averse shortest path interdiction. INFORMS J. Comput. 28 (3), 527–539.

Tanınmış, K., Aras, N., Altınel, İ.K., 2022. Improved x-space algorithm for min–max bilevel problems with an application to misinformation spread in social networks. European J. Oper. Res. 297 (1), 40–52.

Tanınmış, K., Aras, N., Altınel, İ.K., Güney, E., 2020. Minimizing the misinformation spread in social networks. IISE Trans. 52 (8), 850–863.

Tong, H., Prakash, B.A., Eliassi-Rad, T., Faloutsos, M., Faloutsos, C., 2012. Gelling, and melting, large graphs by edge manipulation. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 245–254.

Von Stackelberg, H., Peacock, A., 1952. The Theory of the Market Economy. William Hodge.

Wang, X., Deng, K., Li, J., Yu, J.X., Jensen, C.S., Yang, X., 2018. Targeted influence minimization in social networks. In: Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22. Springer, pp. 689–700.

Wu, H.-H., Küçükyavuz, S., 2018. A two-stage stochastic programming approach for influence maximization in social networks. Comput. Optim. Appl. 69, 563–595.

Yao, Q., Shi, R., Zhou, C., Wang, P., Guo, L., 2015. Topic-aware social influence minimization. In: Proceedings of the 24th International Conference on World Wide Web. pp. 139–140.

Zheng, J., Pan, L., 2018. Least cost rumor community blocking optimization in social networks. In: 2018 Third International Conference on Security of Smart Cities, Industrial Control System and Communications. SSIC, IEEE, pp. 1–5.