Countering Misinformation on Social Networks
Using Graph Alterations

Yigit Ege Bayiz, Ufuk Topcu

Abstract—We restrict the propagation of misinformation in a social-media-like environment while preserving the spread of correct information. We model the environment as a random network of users in which each news item propagates in the network in consecutive cascades. Existing studies suggest that the cascade behaviors of misinformation and correct information are affected differently by user polarization and reflexivity. We show that this difference can be used to alter network dynamics in a way that selectively hinders the spread of misinformation content. To implement these alterations, we introduce an optimization-based probabilistic dropout method that randomly removes connections between users to achieve minimal propagation of misinformation. We use disciplined convex programming to optimize these removal probabilities over a reduced space of possible network alterations. We test the algorithm’s effectiveness using simulated social networks. In our tests, we use both synthetic network structures based on stochastic block models, and natural network structures that are generated using random sampling of a dataset collected from Twitter. The results show that on average the algorithm decreases the cascade size of misinformation content by up to 70% in synthetic network tests and up to 45% in natural network tests while maintaining a branching ratio of at least 1.5 for correct information.

Index Terms—social networks, misinformation, optimization, network design

I. INTRODUCTION

Be it a deliberate spread of controversy caused by a disinformation campaign, or benign misinformation content that cascades through the internet, the propagation of false news is a major issue in social media networks. The increasing public consumption of social media over the last decade has caused more and more people to rely on social media as a source of news [6, 7]. And the attempts to counter misinformation using manual content classification and human moderators have failed to scale up the sheer amount of content [5] that propagates through modern social networks. Therefore over the last decade, automated means of countering false news have drawn great interest.

Existing automated counters to false news mainly focus on the detection of misinformation content. The exact form of these detection algorithms depends on the type of content and the underlying social media network. In general, most misinformation detection methods rely on content classification using some black-box machine learning algorithm that is trained on a dataset labeled by humans. These content classification methods can yield high accuracy. However, these content classifiers still suffer from large biases caused by the biases in the training datasets [7]. And their use in social media platforms can lead to ethnic, religious or political discrimination within the platform.

In this work, we deviate from the existing literature that mainly focuses on misinformation detection, and instead, we approach the problem of countering misinformation as an optimal control problem on a network. In this approach, we study the problem of altering the social network dynamics in a way that restricts misinformation spread while keeping the propagation of true content above some acceptable level. As such, we represent the problem of countering misinformation as an optimization problem on a social network model and then solve this optimization problem to find real-world changes to the social network that reduce the misinformation spread over the network.

In our social network model, we represent the propagation of news on a social media platform using a percolation process model. In this model, each news item propagates in a small-world network of users in consecutive cascades. In each cascade, the users in the network that believe a news content is correct re-share the news content probabilistically. Then, each user that receives the shared news content, either believes in it or discards it based on a probability distribution determined by the polarization of the sharing and receiving users and the reflexivity of the receiving user.

The evidence suggests that there are subtle yet detectable differences in the cascade behaviors of misinformation and true content. User polarization and reflexivity are thought to be the main drivers of this difference [5] [18]. Thus, the above model can capture different propagation patterns of misinformation and true content. Our approach to countering misinformation relies on this difference in propagation patterns to discriminate between different content types. We exploit this propagation difference to design alterations on the network of users that selectively hinders the spread of misinformation containing news while maintaining acceptable propagation of true content.

The specific methods that can be used to control the content flow over the network vary significantly depending on the capabilities and the structure of the underlying online platform. We assume a social media environment that can acquire usage data from the users and can estimate the polarization of the user as well as the probability to re-share a particular news content. Under this assumption, we propose a method called the Dropout Method. This method relies on selectively limiting the content flow over the network by the random omission of news items in a user’s news feed with a predetermined dropout probability. These probabilities must be set up to

Y. E. Bayiz, and U. Topcu are with University of Texas at Austin, TX 78712 USA. E-mails: {egebayiz, utopcu}@utexas.edu.
reduce misinformation spread while minimally affecting the spread of correct content. To achieve this discrimination between misinformation and true content, we let the dropout probabilities depend on the polarization of both the sharing and receiving users. As mentioned before, these quantities are known to affect misinformation and true content flow differently, thus they can be used to identify news shares that are likely to contain misinformation.

A. Main Contributions

We have two main contributions:

1) In section 3, we develop an optimization-based approach to model the problem of countering misinformation through alterations in the social network structure.

2) In section 4, we seek approximate solutions to the problem we develop in section 3, and ultimately develop an algorithm that can counter misinformation through alterations in the social network structure.

B. Related Work

1) Misinformation Propagation: The existing works on misinformation propagation can be separated into two different categories, the first of these categories focuses on finding mathematical models that describe the propagation of misinformation content. Most of these models describe the propagation of misinformation using established epidemiological models. The epidemiological models that have been most widely used in modeling viral content are, susceptible-infected-susceptible (SIS) [8, 11], susceptible-infected-removed (SIR) [22, 25], and susceptible-exposed-infected-removed (SEIR) [15, 24]. All of these models describe misinformation propagation over a social network by classification the users on the social network to different groups and modeling how these groups evolve over time. In their work Raponi et. al. provides a comprehensive analysis of these epidemiological models and their use in modeling misinformation spread [20]. In our work we use an SIR-based model for content propagation as it is a widely accepted model for modelling fake news and it is simple to analyze.

The second category of works on misinformation propagation focuses on discriminating misinformation spread from the spread of other content. In their work Zhao et. al. statistically show that the propagation patterns of fake news differ predictably form other content [26]. Wu et. al. uses support vector machine classifiers to detect misinformation campaigns based on propagation patterns [23]. There are also mixed approaches to misinformation detection that uses both automated content classifiers and propagation patterns to detect misinformation. Varol et. al. uses a supervised learning approach based on k-nearest neighbors classifiers that uses sentiment values and propagation patterns to identify promoted campaigns on social media [21]. Our approach is similar to these works in the sense that we seek discrimination between misinformation and other content. However, our method does not attempt to explicitly identify misinformation.

2) Countering Misinformation: There are some existing works that attempt to limit the propagation of misinformation. In their work Fan et. al. proposes two models for multiple competing diffusion processes on network and investigates the problem of containing rumor spread on a competitive diffusion model [4]. Similarly Litou et. al. model the competition between misinformation and credible information on a network using a novel dynamic linear threshold model and investigate the problem of finding optimal set of users on a network to initiate the propagation of credible content [13, 14]. More recently there has been work on refining this approach by considering location [27] or community [16] structures of the underlying social network. Unlike these works we do not focus on minimizing the influence of misinformation by maximizing the influence of a competing diffusion model. Instead we focus on altering the social media dynamics in a way that passively reduces misinformation spread without requiring any competing credible content.

II. Preliminaries

A. Random Graphs

A random directed graph \( \mathcal{G} = (\mathcal{V}, \{p_{ij}\}) \) composed of a set \( \mathcal{V} \) of vertices, a set \( E \subseteq \mathcal{V} \times \mathcal{V} \) of edges, and a matrix of edge probabilities \( p_{ij} \) that assigns probability to each edge. An instance of a random directed graph \( \mathcal{G} \) is a directed graph \( G = (\mathcal{V}, E') \), where \( E' \subseteq E \) and \( \mathbb{P}((i, j) \in E') = p_{ij} \).

B. Discrete-Time SIR Model on Graphs

A discrete-time susceptible-infected-removed (SIR) model is a contagion propagation model that is often used to model propagation of epidemics. In this model, the contagion spreads in a network over consecutive iterations. Given a random directed graph \( \mathcal{G} = (\mathcal{V}, \{p_{ij}\}) \), at each time \( t \), The SIR model first splits the vertex set \( \mathcal{V} \) into three time-dependent partitions composed of a susceptible set \( S_t \), an infected set \( I_t \), and a removed set \( R_t \). At each time step, each infected node \( i \in I_t \) spreads its infection to all susceptible nodes \( j \in S_t \), with probability \( p_{ij} \). That is,

\[
\mathbb{P}(j \in I_{t+1}|j \in S_t) = 1 - \sum_{i\neq j} (1 - p_{ij}).
\]

Each node that gets infected remains infected for exactly \( m \) turns where \( m \) is a known integer constant. After that, they get removed. That is, for all \( i \in I_t \),

\[
i \in \begin{cases} R_{t+1}, & i \in I_t, \forall t \in \{t-m\ldots t\} \\ L_{t+1}, & \text{otherwise} \end{cases}
\]

In our analysis and for the remainder of this work we take \( m = 1 \). This restricted form of SIR model is equivalent to another commonly used content propagation model called the independent cascade process [10].

In addition to being used extensively in epidemics research, SIR models also see significant use in modelling viral content spread over social networks.
C. Stochastic Block Models

A stochastic block model (SBM) is a random graph model with inbuilt communities. We use the notation \(G_{\text{SBM}}(\mathcal{G}, [b_{uv}]_{uv})\) to refer to an SBM model generated by a finite partition \(\mathcal{G} = \{C_1, C_2, \ldots, C_k\}\) of the set of nodes \(V\) and a \(k \times k\) SBM matrix \([b_{uv}]_{uv}\). We define this SBM as the random graph \(G = (V, E_{ij}[p_{ij}])\), where \(V = \bigcup_{u=1}^{N} C_u\) and the edge probabilities \(p_{ij}\) are given as

\[
p_{ij} = b_{uv} \quad \text{for all} \quad i \in C_u, j \in C_v. \tag{3}
\]

III. MODELS AND PROBLEM SETUP

A. The Social Media Setup

Consider a Twitter-like social media environment that has \(N\) users. We call any user posts or news articles that occur in this social media as content. We use the term true content to describe a content that contains correct or scientific information, and we use the term false content to describe any content which contains misinformation, disinformation or conspiracy.

Contents are spread over the social media environment through shares between users. Once a user receives a content, they can freely choose whether to re-share it again. A content always originates from a subset of users, which we call seeds, and notate as \(I_0\). In practice the number \(|I_0|\) of seeds is almost always small compared to the total number \(N\) of users.

Following the Twitter model, we assume that the information spreads over the network in multicast fashion. That is, whenever a user shares a content, the content is transmitted to all receivers simultaneously. We also assume that each user can share or re-share a particular content only once.

In practice, user shares can occur at any time \(t \in [0, \infty)\). However, modelling and analysis in this continuous time domain is difficult. Thus, in our network model we use the discretized time \(t \in \mathbb{N} = \{0, 1, 2, 3, \ldots\}\). The propagation of content in the network can be summarized as the following iterative process.

1) Set \(t = 0\).
2) Content originates form seeds \(I_0\). The seeds share the content.
3) Set \(t \leftarrow t + 1\).
4) Some of the users that receive the content decide to re-share it again.
5) If there are new re-shares, return to step 3.

B. Modelling Content Propagation

Given a social network with \(N\) users, let \(\mathcal{G} = \{C_1, C_2, \ldots, C_k\}\) denote the partition on the set of users that is induced by user polarization. These partitions can be generated by the political or moral opinions of the users, as well as the echo chambers that exist over the network. We model each user’s probability of re-sharing received content using two quantities:

- \(r_i^+\): The probability that the user \(i\) re-shares a received true content.
- \(r_i^-\): The probability that the user \(i\) re-shares a received false content.

These quantities reflect both reflexivity, which is the ability to discriminate true and false content, and the probability of each user re-sharing the content they believe to be true.

We model the probability of a content shared by user \(i \in C_u\) to be received by user \(j \in C_v\) as a constant probability \(c_{uv}^-\) for false and \(c_{uv}^+\) for true content respectively. In practice, for the echo chambers that result from the user polarizations encourage content sharing between agents within the same polarization class and discourage content propagation across different polarization classes. We can write the total probability of content being transferred from user \(i \in C_u\) to \(j \in C_v\) as

\[
p_{ij}^+ = r_i^+ c_{uv}^+ \quad \text{for false content}, \tag{4a} \\
p_{ij}^- = r_i^- c_{uv}^- \quad \text{for true content}. \tag{4b}
\]

Following the social media setup in section 3.1, we model the content propagation using an SIR model with an infectious period of 1. Here the partition \(S_t\) represents the users that have not yet received a piece of content at time \(t\), \(I_t\) represents the users that have received the content in the current time step \(t\), and \(R_t = \bigcup_{s=t}^{\infty} I_s\) is the set of nodes that have previously received the content.

We can approximate these content propagation dynamics as an SIR model with infectious period of 1 on one of two stochastic block models:

\[
\mathcal{G}^- := G_{\text{SBM}}(\mathcal{G}, [b_{uv}]_{uv}) \quad \text{for false content}, \tag{5a} \\
\mathcal{G}^+ := G_{\text{SBM}}(\mathcal{G}, [b_{uv}]^+_{uv}) \quad \text{for true content}, \tag{5b}
\]

where for all \(u, v \in \{1, 2, \ldots, k\}\) we have

\[
b_{uv}^+ = \frac{1}{|C_u|} \sum_{i \in C_u} r_i^+ c_{uv}^+, \tag{6a} \\
b_{uv}^- = \frac{1}{|C_u|} \sum_{i \in C_u} r_i^- c_{uv}^-. \tag{6b}
\]

The difference between the stochastic block models \(\mathcal{G}^+\) and \(\mathcal{G}^-\) results in different content propagation characteristics to be predicted by the SIR model. This reflects the difference in content propagation patterns that can be seen between real-world true and false content. When correctly fitted to the actual social media network, these simplified content propagation models are often capable of sufficiently capturing the difference in propagation patterns between true and false content.

We assume throughout this work that we know \(b_{uv}^-\) and \(b_{uv}^+\). This is a reasonable assumption since we can fit these SBMs to the real data collected from social media by directly estimating \(b_{uv}^-\) and \(b_{uv}^+\). A simple method of doing this is by first observing the propagation patterns of sample contents which are known to be either true or false, and then using a frequentist estimation of \(b_{uv}^-\) and \(b_{uv}^+\) from the observed propagation patterns. As more propagation data on true/false contents become available, this estimation can be repeated periodically to refine the estimates for \(b_{uv}^-\) and \(b_{uv}^+\) over time.

This estimation process requires reliable knowledge of whether the observed sample contents are true or false. Therefore we require reliable content classification to fit the SBMs.
$\mathcal{G}^+$ and $\mathcal{G}^-$ to the actual social media. To minimize the bias and fairness issues associated with automated content classification systems, we suggest doing the content classification either using user responses to the content (likes, comments, etc.) or manually by expert human moderators. This is possible, since there is no strict requirement to classify content during its propagation period, and the classification can easily be done afterward without any constraint on time. As we elaborate in the following sections, the fact that automated content classification is superfluous for estimating content propagation models is one of the major advantages of our approach.

C. Graph Alterations and Dropouts

To counter the spread of false content over the social media network we need to determine how we can control the content propagation over the network. The classical way of achieving this is first determining if a piece of content is true or false using automated algorithms, and then restricting, or banning the content which is determined to be false. This is an effective means of stopping the propagation of content that is classified as false. However, as mentioned previously, this approach suffers from its explicit reliance on automated content classification methods, and the issues this reliance brings.

To resolve this reliance on automated agents, we introduce a network-design-based approach to counter false content called graph alterations. Let $f: V \times V \times [0, 1] \to [0, 1]$ be a function that given a node pair $i,j$ and content transfer probability $p_{ij}$, generates an altered content transfer probability of $f(i,j,p_{ij})$. We define a graph alteration $\mathcal{G}_f: \mathcal{G} \mapsto \tilde{\mathcal{G}}$ as the mapping induced by $f$ between two random graphs. That is for any random graph $\mathcal{G} := (V,\{p_{ij}\}_{ij})$ we have

$$\mathcal{G}_f(\mathcal{G}) := \tilde{\mathcal{G}} := (V,\{f(i,j,p_{ij})\}_{ij}).$$

(7)

In other words, given a random graph $\mathcal{G}$, with transfer probabilities $\{p_{ij}\}_{ij}$ $\mathcal{G}_f(\mathcal{G})$ is a random graph with altered transfer probabilities $\{f(i,j,p_{ij})\}_{ij}$.

Suppose that we have two random graphs $\mathcal{G}^+$ and $\mathcal{G}^-$ describing the propagation of true and false content respectively on a social network. For any fixed $f$, applying the same graph alteration $\mathcal{G}_f$ to $\mathcal{G}^-$ and $\mathcal{G}^+$ provides a method to alter the structure of both of these graphs simultaneously in a way that does not explicitly depend on the content type. Each graph alteration $\mathcal{G}_f$ corresponds to a change in the social network structure that results in the true and false contents to propagate according to random graphs $\mathcal{G}_f(\mathcal{G}^+)$ and $\mathcal{G}_f(\mathcal{G}^-)$ respectively.

The set of graph alterations that are feasible to implement on a social media network depend heavily on the design and capabilities of the social media platform. In this work, we focus on the graph alterations corresponding to randomized content dropouts. We define a dropout as the artificial prevention of a content transfer between two agents. For example, in a Twitter-like social media platform, we can implement such randomized content dropouts by artificially excluding a content from the receiver’s feed. In the SBM model described in Section 3.4, we model a random dropout between two users $i \in C_u$ and $j \in C_v$ using an altered content transfer probability function.

$$f_d(i,j,p_{ij}) := d_{uv}p_{ij},$$

(8)

where $d_{uv}$ are the dropout probabilities. Then under graph alteration $\mathcal{G}_f$, the altered SBMs corresponding to true and false content propagation graphs are

$$\mathcal{G}^+ := \mathcal{G}_f(\mathcal{G}^+) := \mathcal{G}_{SBM}(\mathcal{G}, [\tilde{b}_{uv}]_{uv})$$

for false content,

$$\mathcal{G}^+ := \mathcal{G}_f(\mathcal{G}^+) := \mathcal{G}_{SBM}(\mathcal{G}, [\tilde{b}_{uv}]_{uv})$$

for true content, (9b)

where

$$\tilde{b}_{uv}^+ := d_{uv}b_{uv}^+,$$

(10a)

$$\tilde{b}_{uv}^- := d_{uv}b_{uv}^-.$$  

(10b)

Assuming that the original SBMs $\mathcal{G}^-$ and $\mathcal{G}^+$ are accurate models of content propagation over the network, we can use these altered SBMs to predict the effect of graph alterations on the real-world social network.

D. Problem Statement

In the most general case, the problem of countering misinformation is finding a graph alteration $\mathcal{G}_f$ that minimizes the predicted propagation of false content, whilst keeping the predicted propagation of true content above some acceptable level. That is for a given safety parameter $\alpha$, at every time $t$ we wish to solve the following optimization problem.

$$\begin{aligned}
\min_{\mathcal{G}_f} & \quad E_{\mathcal{G}_f} \left[ |I_{t+1}| \right] \left| S_t, I_t, R_t \right|, \\
\text{s.t.} & \quad E_{\mathcal{G}_f} \left[ |I_{t+1}| \right] \left| S_t, I_t, R_t \right| \geq \alpha |I_t|.
\end{aligned}$$

(11)

The general case given in (11) is a non-convex optimization problem over all possible graph alterations $\mathcal{G}_f$. This is intractable for large networks. However, we can reduce it into a simpler problem by restricting and parametrizing the graph alterations $\mathcal{G}_f$. Throughout the rest of this work, we restrict our analysis to graph alterations $\mathcal{G}_f$ that are generated by randomized dropouts $[d_{uv}]_{uv}$ where $d_{uv}$ is the dropout probability of a content transfer from a user in polarization class $C_u$ to a user in polarization class $C_v$. This yields the following optimization problem:

$$\begin{aligned}
\min_{d \in [0,1][k,k]} & \quad E_{\mathcal{G}_f} \left[ |I_{t+1}| \right] \left| S_t, I_t, R_t \right|, \\
\text{s.t.} & \quad E_{\mathcal{G}_f} \left[ |I_{t+1}| \right] \left| S_t, I_t, R_t \right| \geq \alpha |I_t|.
\end{aligned}$$

(12)

where $f_d$ is the altered transfer probability function defined in eq. (8), and $\mathcal{G}_f$ is the graph alteration induced by $f_d$.

After solving these optimization problems with model social networks $\mathcal{G}^-$ and $\mathcal{G}^+$, we use the optimal graph alterations found by these problems to alter the content transfer probabilities of the real-world social network. Since the optimal solutions of problems (12) and (11) are time-dependent, we need to update the graph alteration at each time $t$ by re-evaluating the optimal solution to the optimization problem at hand based on the observed $S_t, I_t, R_t$. 

As a general solution template, we consider the dynamic false content minimization loop given in Algorithm 1. We run Algorithm 1 independently for each content that propagates over the network. In Algorithm 1, $\text{OPT}(\text{problem } (12))$ refers to the optimal solution of problem (12), and observe $(S_{t+1}, I_{t+1}, R_{t+1} | S_t, I_t, R_t, \mathcal{G})$ returns the observed $S_{t+1}, I_{t+1}, R_{t+1}$ sets generated by an SIR model on an altered social network $\mathcal{G}$ with known current state $S_t, I_t, R_t$. That is, the $S_{t+1}, I_{t+1}, R_{t+1}$ denotes the next set of susceptible, infected, removed users given that the real social network is altered using a dropout $d^*$. This dropout is the optimal dropout based on our model networks $\mathcal{G}^+, \mathcal{G}^-$, which are stochastic block models as described in the previous section. Intuitively, given a piece of content that propagates as an SIR model on a real-world social network $\mathcal{G}$ with unknown transfer probabilities, Algorithm 1 attempts to minimize the spread of the content if it propagates like a false content, and preserves the spread of the content if it propagates like a true content.

Algorithm 1 False Content Minimization

Require: Model Networks $\mathcal{G}^+, \mathcal{G}^-$, Real-World Network $\mathcal{G}$, Set of seed users $I_0$, Safety parameter $\alpha$.

1: $t \leftarrow 0$
2: $S_0 \leftarrow V \setminus I_0$
3: $R_0 \leftarrow \emptyset$
4: while $|I_t| > 0$ do
5: $d^* \leftarrow \text{OPT}(\text{problem } (12))$
6: $\mathcal{G} \leftarrow \mathfrak{A}_{d^*}(\mathcal{G})$ \{ Alter the real-world social network. \}
7: $S_{t+1}, I_{t+1}, R_{t+1} \leftarrow \text{Observe}(S_{t+1}, I_{t+1}, R_{t+1} | S_t, I_t, R_t, \mathcal{G})$
8: $t \leftarrow t + 1$
9: end while

The problem given in eq. (12) is a non-convex problem. To solve it, we formulate an asymptotic approximation of it by considering the behavior of $\mathcal{G}^+ = \mathfrak{A}_{d^*}(\mathcal{G}^+)$ $[|I_{t+1}| | S_t, I_t, R_t]$ as the number $N$ of users diverges towards infinity. This is a reasonable approximation since the number $N$ of users in real social networks is often large enough that the inaccuracies caused by the asymptotic approximation is negligible compared to other sources of model inaccuracy.

We define $I^+_u = I_u \cap C_u$ as the set of infected users in polarization class $C_u$ at iteration $t$. Similarly, we also let $R^+_u = R_u \cap C_u$, and $S^+_u = S_u \cap C_u$. Then, for any $j \in C_u$ we have

$$\mathbb{P}_{\mathcal{G}^+}[j \in I_{t+1} | S_t, I_t, R_t] = \begin{cases} 1 - \prod_{u=1}^{k} (1 - d_{uv}b_{uv}^+), & j \in S_t, \\ 0, & j \notin S_t \end{cases}$$

This probability immediately follows from the transition probabilities of the altered SBM $\mathcal{G}^+ = \mathfrak{A}_{d^*}(\mathcal{G}^+)$ corresponding to the false content.

The social networks that we are interested in often have a large number of users. Thus, we are interested in the asymptotic behavior of eq. (13) for large $N$. For such large networks we can approximate (13) as

$$\mathbb{P}_{\mathcal{G}^+}[j \in I_{t+1} | S_t, I_t, R_t] = \begin{cases} 1 - \prod_{u=1}^{k} \exp(-|I_u|^d b_{uv}^+), & j \in S_t, \\ 0, & j \notin S_t \end{cases}$$

Under this approximation, for true content we have

$$\mathbb{E}_{\mathcal{G}^+}[|I_{t+1}| | S_t, I_t] = \sum_{v=1}^{k} |S^+_v| \left( 1 - \exp \left( - \sum_{u=1}^{k} |I_u|^d b_{uv}^+ \right) \right),$$

and similarly, for false content, we write

$$\mathbb{E}_{\mathcal{G}^-}[|I_{t+1}| | S_t, I_t] = \sum_{v=1}^{k} |S^+_v| \left( 1 - \exp \left( - \sum_{u=1}^{k} |I_u|^d b_{uv}^+ \right) \right).$$

Using equations (15) and (16), we can rewrite the optimization problem given in (12) as

$$\min_{d \in [0,1]^k} \sum_{v=1}^{k} |S^+_v| \left( 1 - \exp \left( - \sum_{u=1}^{k} |I_u|^d b_{uv}^+ \right) \right) \geq \alpha |I_t|. \quad (17a)$$

There is no guarantee that the optimization problem in eq. (17) is feasible. In fact, since the left hand side of the constraint (17) is monotonically increasing with $d_{uv}$ for all $u, v \in \{1, \ldots, k\}$ the problem (17) is feasible if and only if we have

$$\sum_{v=1}^{k} |S^+_v| \left( 1 - \exp \left( - \sum_{u=1}^{k} |I_u|^d b_{uv}^+ \right) \right) \geq \alpha |I_t|. \quad (17b)$$

That is, the problem (17) is feasible exactly when $(\lambda t)$ has branching ratio greater than or equal to $\alpha$ in the SIR model defined on the non-altered true content graph $\mathcal{G}^+$. Therefore choosing $\alpha$ too large can lead to infeasibility.

Another issue to note is that due to the dynamics of the SIR model given in equations (1) and (2), whenever we have $|I_t| = 0$ for some $\tau$, we guarantee $|I_t| = 0$ for all $t > \tau$. This means that even though (17) might be feasible, the propagation of true content can halt if a random event leads to $|I_t| = 0$ at some $\tau$. Clearly, the probability of this event $|I_{\tau}| = 0$ decreases with larger safety parameter $\alpha$. But as stated previously, too large of a choice for the parameter $\alpha$ leads to infeasibility in the problem (17). Therefore, when choosing a safety parameter $\alpha$ one needs to consider a trade-off between feasibility, and robustness to probabilistic effects. In Lemma 1 we investigate this trade-off further and characterize the relation between $\alpha$ and the probability that the $|I_t| = 0$ given that the problem given in (17) is feasible.

**Lemma 1.** Suppose that there exists some $T \in \mathbb{N}$ such that the optimization problem (12) is feasible for all $t \in [0, T]$, and let $(S_t, I_t, R_t)$ be the stochastic SIR process generated as in Algorithm 1. Then if $\mathcal{G} = \mathcal{G}^+$, then for any non-negative $\lambda$ we have

$$\mathbb{P}_{\mathcal{G}}[\text{inf}_{0 \leq t \leq T} |I_t| = 0] \leq \mathbb{E} \left[ e^{-\lambda |I_{\tau}|} \right] = M_{|I_{\tau}|} \left( -\frac{\lambda}{\alpha^2} \right),$$

where $M_{|I_{\tau}|}$ denotes the moment generating function of $|I_{\tau}|$. 
Proof. Let \((\mathcal{F}_t)_{0 \leq t \leq T}\) be the natural filtration generated by the stochastic process \((|U_t|)_{0 \leq t \leq T}\). Let \(Y_t = \frac{|Y_t|}{\mathbb{E}[Y_t]}\). Then by the constraint of problem (12), for all \(t \in (0, T]\), we have \(\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] = Y_t\). That is, \((Y_t)\) is Martingale. Then by Jensen’s inequality, for all non-negative \(\alpha\) and \(t \in (0, T]\), we have
\[
\mathbb{E}[e^{-\alpha Y_t} \mid \mathcal{F}_t] \geq e^{-\alpha Y_t}.
\]
Notice that \((\mathcal{F}_t)\) is also a natural filtration for the stochastic process \((e^{-\alpha Y_t})\), since \(e^{-\alpha Y_t}\) relates bijectively to \(|U_t|\). Therefore \((e^{-\alpha Y_t})\) is a sub-Martingale sequence. Then,
\[
\mathbb{E}[\inf_{0 \leq t \leq T} |U_t| = 0] = \mathbb{E}[\inf_{0 \leq t \leq T} |Y_t| = 0] \quad (20a)
\]
\[
= \mathbb{E}[\sup_{0 \leq t \leq T} e^{-\alpha Y_t} \geq 1] \quad (20b)
\]
\[
\leq \mathbb{E}[e^{-\alpha Y_T}], \quad (20c)
\]
where the last line follows from Doob’s Martingale inequality. 

The optimization problem (17) can be solved using gradient-based methods. However, it is also possible to simplify it further. We are mainly interested in the initial period of the viral spread of the content. That is, we want to counter false content before it spreads to a significant fraction of users. Similarly for true content, if we can ensure that the spread of true content is not restricted in the first couple of iterations, it is likely that a significant fraction of users will eventually receive the true content. Thus in practical applications, \(N\) is usually much larger than \(I_t\). Under this assumption, the inequalities
\[
\exp \left( \sum_{u=1}^{k} |l_t^u| d_{uv} b_{uv}^+ \right) \geq 1 - \sum_{u=1}^{k} |l_t^u| d_{uv} b_{uv}^-,
\]
\[
\exp \left( \sum_{u=1}^{k} |l_t^u| d_{uv} b_{uv}^- \right) \geq 1 - \sum_{u=1}^{k} |l_t^u| d_{uv} b_{uv}^-,
\]
becomes tight. Therefore in the regime \(N > I_t\), we can approximate the solution the optimization problem (17) using the following optimization problem,
\[
\min_{d \in [0,1]^{k \times k}} \sum_{v=1}^{k} \sum_{u=1}^{k} |S_v^u||l_t^u| d_{uv} b_{uv}^- + \lambda |S_v^u||l_t^u| d_{uv} b_{uv}^+ \quad (22a)
\]
\[
s.t. \quad \sum_{v=1}^{k} \sum_{u=1}^{k} |S_v^u||l_t^u| d_{uv} b_{uv}^+ \geq \alpha |I_t|. \quad (22b)
\]
The above form is a linear program and it can be solved very efficiently using existing linear program solvers. As before, the optimization problem (22) is feasible if and only if we have
\[
\sum_{v=1}^{k} \sum_{u=1}^{k} |S_v^u||l_t^u| b_{uv}^+ \geq \alpha |I_t|. \quad (23)
\]
In application, viral true content almost always satisfies this feasibility condition for some \(\alpha \geq 1\) in the initial propagation period. However, less viral content may violate the feasibility condition. In this case, our original optimization goals cannot be reached. To counter this issue, for content that violates the feasibility criterion (23) infeasible we can soften the linear program (22). That is,
\[
\min_{d \in [0,1]^{k \times k}} \sum_{v=1}^{k} \sum_{u=1}^{k} |S_v^u||l_t^u| d_{uv} b_{uv}^- + \lambda |S_v^u||l_t^u| d_{uv} b_{uv}^+ \quad (24)
\]
where \(\lambda\) is a weight parameter that signifies the importance of preserving true content relative to the importance of suppressing false content.

We can use the linear programs (22) and (24) and in conjunction to provide a general approximate solution the our main problem (12) given in section 3.4. The Algorithm 2 provides a method to achieve which this conjunction. Here \(OPT(\ast)\) refers to the optimal solution on the optimization problem *, which in the case of Algorithm 2 can be found efficiently using existing linear program solution methods.

**Algorithm 2** False Content Minimization Using Dropouts

**Require:** Networks \(\mathcal{G}^-, \mathcal{G}^+\), Real-World Network \(\mathcal{G}\), Set of seed users \(I_0\), Safety parameter \(\alpha\), Weight parameter \(\lambda\).

1: \(t \leftarrow 0\)
2: \(S_0 \leftarrow V \setminus I_0\)
3: \(R_0 \leftarrow \emptyset\)
4: while \(|I_t| > 0\) do
5: \(\hat{S}_t \leftarrow \mathcal{Af}^{\mathcal{G}}(\mathcal{G}) \{\text{Alter the real-world social network.}\}\)
6: \(d^* = \text{OPT}(\text{linear program (22)})\)
7: else
8: \(d^* = \text{OPT}(\text{linear program (24)})\)
9: end if
10: \(\langle I_t+1, R_t+1 \rangle \leftarrow \text{Observe} \langle S_t+1, I_t+1, R_t+1 \rangle \}
11: \(S_t+1, I_t+1, R_t+1 \leftarrow \langle S_t+1, I_t+1, R_t+1 \rangle \}
12: \(t \leftarrow t + 1\)
13: end while

**V. Experimental Results**

We test Algorithm 2 both on synthetic stochastic block model networks and on a real misinformation dataset collected over Twitter. The dataset we used is called WICO [19] which contains over 3500 separate tweets and status updates collected between January 2020 and July 2020. In these tests, we use the total cascade size \(R_m\), which is the total size of the set of users that have received a piece of content after the SIR propagation terminates, as the performance metric. For the case of true content, we want \(R_m\) to be high, and for the case of false content, we want \(R_m\) to be low.

**A. Experiments Using Synthetic Model**

In this section, we test the performance of Algorithm 2 using synthetic social networks that are modeled as SBMs. We test the effectiveness of Algorithm 2 over four different test configurations. We name these configurations as follows: Balanced with 2 partitions, unbalanced with 2 partitions, balanced with 3 partitions, and unbalanced with 3 partitions. All of these configurations have 1000 users. The balanced configurations have partition sizes of [500, 500] for 2 partition case and [334, 333, 333] for 3 partition case. The unbalanced
the SBM matrices for probabilities for true and false content respectively. We define
we define the SBMs $G$ and for false content we define in each dimension. For each $x \in \mathbb{R}^n$ of algorithm 2, we select two parameters different content propagation dynamics on the performance true and false content propagation. To test the effect of base matrix as,

Similarly, for configurations with 3 partitions, we define this base matrix as,

$$b_{\text{base}} := \begin{bmatrix} 0.01 & 0.002 \\ 0.002 & 0.01 \\ 0.002 & 0.002 & 0.01 \end{bmatrix}.$$ \hfill (25)

We use these base matrices to generate SBMs that simulate true and false content propagation. To test the effect of different content propagation dynamics on the performance of algorithm 2, we select two parameters $x \in [0,0.01]$ and $y \in [0,0.001]$. For each configuration, we sweep across the range of possible $x$ and $y$ combinations with 50 subdivisions in each dimension. For each $x$ and $y$ choice, we generate the SBMs $\mathcal{G}^+$ and $\mathcal{G}^-$ that describe the content transfer probabilities for true and false content respectively. We define the SBM matrices for $\mathcal{G}^+$ and $\mathcal{G}^-$ as follows: For true content we define

$$[b^+]_{xy} = b_{\text{base}} + xI - y(J - I),$$ \hfill (27)

and for false content we define

$$[b^-]_{xy} = b_{\text{base}} - xI + y(J - I),$$ \hfill (28)

where $I$ is the identity matrix and $J$ is the all-ones matrix. Then we simulate content propagation over these networks and determine the cascade sizes $R_\infty$ that result from different choices for $\alpha$ and $\lambda$.

Table I summarizes the normalized mean cascade size $E[R_\infty]/N$ and ratio of tests that have cascade size less than $N/10$ to all tests. These statistics are collated over the complete range of all $x$ and $y$ combinations. For all configurations, we test two different parameter assignments for $\alpha$ and $\lambda$. The rows indicated as $(\alpha, \lambda) = (-, -)$ are control groups with no network alterations. This table shows that regardless of the choice of parameters $\alpha$ and $\lambda$, on average Algorithm 2 manages to reduce false content more than it reduces true content. Moreover, $\mathbb{P}[R_\infty < N/10]$ are much higher on false content compared to true content. The fact that $\mathbb{P}[R_\infty < N/10]$ are high on false content indicates that the cascade size $R_\infty$ has high variance. That is, the performance of Algorithm 2 varies greatly depending on the dynamics of the social network structure.

Figure 1 shows the average cascade size three configurations over the full span of $x,y$ combinations, where the average is computed over 50 separate trials. This figure shows that the Algorithm 2 affects the true content in a similar way across all $x,y$ values. This is of course expected since the constraint in the linear program (22) in Algorithm 2 fixes the expected propagation rate of true content. On the contrary, the effect of Algorithm 2 on the false content depends heavily on both the value of $x$ and $y$ and the overall structure of the social media network. In general, there is a sharp boundary transition
in the cascade size of the false content and the Algorithm 2 tends to either reduce the cascade size of false content to near 0, or have very little impact to the propagation of the false content. This explains the high $P[R_n < N/10]$ values seen in Table I. The sharp transition in cascade size seen in Figures 1b and 1c is caused by the same mechanism that causes the state-transition-like behavior that is present in most complex real networks, where a giant connected component can appear suddenly as we increase the overall expected degree of nodes in a large graph.

### B. Experiments Using Real World Data

We use a pre-existing dataset named WICO [19] for these tests. This dataset contains share times and propagation networks for separate pieces of content. These content are labeled as follows:

1. 5G-Corona Conspiracy: Conspiracy content that claims there is a causation between the Covid-19 pandemic and 5G.
2. Other Conspiracy,
3. Non-Conspiracy.

In this dataset, we re-label content that is labeled as “5G-Corona Conspiracy” or “Other Conspiracy” as false content, and we re-label Non-Conspiracy content as true content.

We then assign polarizations to each user by running modularity based clustering [11] with resolution [12] set to 2 on the union of all graphs in the WICO dataset. The resulting network has 153779 nodes and 216848 edges. The modularity class assignment has 63914 vertices. We restrict the number of partitions by merging all partitions with a number of users less than 1% of the total number of users in the merged graph. The resulting partitioning of the graph has 13 partitions. These partitions correspond to different echo chambers in the network, therefore they are a close approximation for the polarization classes of the users in the social network.

After determining polarization classes, we fit the model networks $\mathcal{G}^+$ and $\mathcal{G}^-$ to the dataset by frequentist estimation of SBM matrices $[h_{uv}]$ and $[h_{uv}]_a$ by counting the number of content transfers between different polarization groups. We then use Algorithm 2 to generate dropout-based alterations on the actual social media network. We simulate the content propagation under these dropout-based alterations by sampling a random content from the dataset and then following its propagation while randomly dropping content transfers based on dropout probabilities $d^*$ given by Algorithm 2.

We test the three different parameter settings for Algorithm 2. These settings are $(\alpha, \lambda) = (1.5, 1)$, $(\alpha, \lambda) = (2, 1.5)$, $(\alpha, \lambda) = (3, 2)$. Table II shows the resulting cascade size statistics, averaged over 500 samples, for these each of these settings as well as a control group which is denoted as $(\alpha, \lambda) = (-, -)$. Contrary to the previous synthetic tests, we do not normalize the expected cascade size value in Table II since the cascade sizes of these networks are very small compared to the total number of users in the network. The fact that the cascade sizes are small is not surprising, since often in the real world social media network only a small fraction of users tend to participate in re-sharing a piece of content they receive due to the vastness of the number of available content and the variability of interests of users. The average performance of Algorithm 2 decreases in these real-world datasets compared to synthetic model test due to the inaccuracies in the SBM models $\mathcal{G}^+$ and $\mathcal{G}^-$. However, for all choices of $(\alpha, \lambda)$ Algorithm 2 achieves discrimination between true and false content.

### Table II: Summary of Results for WICO Dataset Tests

| $\alpha$ | $\lambda$ | $E[R_n]$ True C. False C. | $P[R_n < 5]$ True C. False C. |
|----------|----------|--------------------------|--------------------------|
| -        | -        | 50.1 48.7                | 0.01 0.00                |
| 1.5      | 1        | 32.8 26.1                | 0.05 0.13                |
| 2        | 1.5      | 39.4 28.2                | 0.04 0.09                |
| 3        | 2        | 41.1 36.0                | 0.00 0.02                |

### VI. Conclusion and Future Work

We demonstrated that it is possible to counter misinformation without explicit identification of misinformation content by altering the content propagation dynamics on social networks. A major advantage of our approach is that it does not require the system to be able to identify if a particular news item contains misinformation content or not. Furthermore, our approach can be used in conjunction with these detection algorithms to improve the effectiveness of misinformation control while maintaining some of the advantages offered by the network-design-based approach. In our future studies, we will investigate this possibility further.
Throughout this work we have assumed that content is either true and false. In reality this clear of a separation between true and false content types is rarely possible. It is possible to extend our methods and algorithms to admit more content types, which can increase the performance of Algorithm 1 since more content types can achieve a more nuanced description of real misinformation dynamics. However, this extension requires modifications on the problem statement and the formulation of the problem.

ACKNOWLEDGEMENTS

This work was supported in part by ONR N00014-21-1-2502 and NSF 1652113.

REFERENCES

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* (2008), P10008.

[2] Béla Bollobás. 2001. *Random Graphs* (2 ed.). Cambridge University Press. [https://doi.org/10.1017/CBO9780511814068](https://doi.org/10.1017/CBO9780511814068)

[3] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559. [https://doi.org/10.1073/pnas.1517441113](https://doi.org/10.1073/pnas.1517441113)

[4] Lidan Fan, WeiLi Wu, Xuming Zhai, Kai Xing, Wonjun Lee, and Ding-Zhu Du. 2014. Maximizing rumor containment in social networks with constrained time. *Social Network Analysis and Mining* 4, 1 (2014), 1–10.

[5] Eric Gilbert. 2013. Widespread Underprovision on Reddit. In *ACM Conference on Computer Supported Cooperative Work*. [https://doi.org/10.1145/2441776.2441866](https://doi.org/10.1145/2441776.2441866)

[6] Alison L Hill, David G Rand, Martin A Nowak, and Nicholas A Christakis. 2010. Infectious disease modeling of social contagion in networks. *PLOS computational biology* 6, 1 (2010).

[7] Vaishali Vaibhav Hirlekar and Arun Kumar. 2020. Natural Language Processing based Online Fake News Detection Challenges – A Detailed Review. In *International Conference on Communication and Electronics Systems*. [https://doi.org/10.1109/ICCES48766.2020.9137915](https://doi.org/10.1109/ICCES48766.2020.9137915)

[8] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of Workshop on Social Network Mining and Analysis*. [https://doi.org/10.1080/17512786.2017.1423237](https://doi.org/10.1080/17512786.2017.1423237)

[9] Edson C. Tandoc Jr. 2019. Tell Me Who Your Sources Are. *Journalism Practice* 13, 2 (2019), 178–190. [https://doi.org/10.1080/17512786.2017.1423237](https://doi.org/10.1080/17512786.2017.1423237)

[10] David Kempe, Jon Kleinberg, and Eva Tardos. 2003. Maximizing the Spread of Influence through a Social Network. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. [https://doi.org/10.1145/956750.956769](https://doi.org/10.1145/956750.956769)

[11] Masahiro Kimura, Kazumi Saito, and Hiroshi Motoda. 2009. Efficient estimation of influence functions for SIS model on social networks. In *International Joint Conference on Artificial Intelligence*.

[12] Renaud Lambiotte, Jean-Charles Delvenne, and Maurizio Barahona. 2008. Laplacian Dynamics and Multiscale Modular Structure in Networks. *arXiv* (2008).

[13] Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. 2016. Real-Time and Cost-Effective Limitation of Misinformation Propagation. In *IEEE International Conference on Mobile Data Management*. [https://doi.org/10.1109/MDM.2016.33](https://doi.org/10.1109/MDM.2016.33)

[14] Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. 2017. Efficient and timely misinformation blocking under varying cost constraints. *Online Social Networks and Media* 2 (2017), 19–31.

[15] Qiming Liu, Tao Li, and Meici Sun. 2017. The analysis of an SEIR rumor propagation model on heterogeneous network. *Physica A: Statistical Mechanics and its Applications* 469 (2017), 372–380.

[16] Jiaguo Lv, Bin Yang, Zhen Yang, and Wei Zhang. 2019. A community-based algorithm for influence blocking maximization in social networks. *Cluster Computing* 22, 3 (2019), 5587–5602.

[17] Regina Marchi. 2012. With Facebook, Blogs, and Fake News, Teens Reject Journalistic “Objectivity”. *Journal of Communication Inquiry* 36, 3 (2012), 246–262. [https://doi.org/10.1177/0196859912458700](https://doi.org/10.1177/0196859912458700)

[18] Gordon Pennycook and David G. Rand. 2021. The Psychology of Fake News. *Trends in Cognitive Sciences* 25, 5 (2021), 388–402. [https://doi.org/10.1016/j.tics.2021.02.007](https://doi.org/10.1016/j.tics.2021.02.007)

[19] Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. In *ACM Workshop on Open Challenges in Online Social Networks*. [https://doi.org/10.1145/3472720.3483617](https://doi.org/10.1145/3472720.3483617)

[20] Simone Raponi, Zeinab Khalifa, Gabriele Oligeri, and Roberto Di Pietro. 2022. Fake news propagation: a review of epidemic models, datasets, and insights. *ACM Transactions on the Web* 16, 3 (2022), 1–34.

[21] Onur Varol, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2017. Early detection of promoted campaigns on social media. *EPJ Data Science*, 1–19.

[22] Ya-Qi Wang and Jing Wang. 2017. SIR rumor spreading model considering the effect of difference in nodes’ identification capabilities. *International Journal of Modern Physics* 28, 05 (2017).

[23] Ke Wu, Song Yang, and Kenny Q. Zhu. 2015. False rumors detection on Sina Weibo by propagation structures. In *IEEE International Conference on Data Engineering*. [https://doi.org/10.1109/ICDE.2015.7113322](https://doi.org/10.1109/ICDE.2015.7113322)

[24] Ling-Ling Xia, Guo-Ping Jiang, Bo Song, and Yu-Rong Song. 2015. Rumor spreading model considering hesitating mechanism in complex social networks. *Physica A: Statistical Mechanics and its Applications* 437 (2015), 295–303.
[25] Laijun Zhao, Hongxin Cui, Xiaoyan Qiu, Xiaoli Wang, and Jiajia Wang. 2013. SIR rumor spreading model in the new media age. *Physica A: Statistical Mechanics and its Applications* 392, 4 (2013), 995–1003.

[26] Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin. 2020. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science* 9, 1 (2020), 7. [https://doi.org/10.1140/epjds/s13688-020-00224-z](https://doi.org/10.1140/epjds/s13688-020-00224-z)

[27] Wenlong Zhu, Wu Yang, Shichang Xuan, Dapeng Man, Wei Wang, Xiaojiang Du, and Mohsen Guizani. 2019. Location-based seeds selection for influence blocking maximization in social networks. *IEEE Access* (2019), 27272–27287.