Abstract

We study how communication platforms can improve social learning without censoring or fact-checking messages, when they have members who deliberately and/or inadvertently distort information. We analyze message fidelity as a function of social network depth (how many times information is relayed) and breadth (the number of relay chains accessed). Message fidelity can be improved by capping depth or, if that is not possible, limiting breadth; e.g., by capping the number of people to whom someone can forward a given message. Although they reduce total communication, such caps increase the fraction of received messages originating closer to the receiver, thereby increasing the signal to noise ratio. We characterize how the expected number of true minus false messages depends on breadth and depth of the network and the noise structure.

JEL Classification Codes: D83, D85, L14, O12, Z13

Keywords: Social Learning, Communication, Noise, Mutation, Bias, Fake News, Censorship, Misinformation, Disinformation
The safety of our democracy is more important than shareholder dividends and CEO salaries ... That's why I'm calling on [tech companies] to take real steps right now to fight disinformation.

Elizabeth Warren

I don't think that Facebook or Internet platforms in general should be arbiters of truth.

Mark Zuckerberg

1 Introduction

Misinformation has always been a social concern, but there has recently been a resurgence of interest in addressing it, both because social media and messaging platforms make (mis)information spreading easier (e.g., see Vosoughi et al. (2018)) and because they have made its impact more transparent – with wide-ranging implications from voter attitudes toward election outcomes to their views on vaccination. The most common approach to regulate communication on online platforms involves some form of fact-checking and flagging or censorship. However, policing communication is challenging and problematic, due to the enormous volume and fast pace of online communication and the potential for bias by whoever is “determining the truth,” be it a government or private enterprise, an algorithm or crowd. Moreover, communication on some platforms is intentionally encrypted to protect user privacy, making fact-checking impossible.

Given these challenges, we study policies that improve informational content and social learning, without relying on anyone to know or decide what is true. In particular, we study how people learn as a function of their networks when information is subject to mutation, deliberate manipulation, and transmission failure. We characterize how learning depends on the depth and breadth of a person’s network. Increasing depth and breadth of a network increases the number of messages passed and therefore received. But it also increases the number of distant sources, from which surviving messages are less accurate, more than it increases the number of nearby sources. Our analysis shows how and when limiting the network can improve the accuracy of overall content without the need for censorship or message monitoring. Our results show how one can satisfy Warren’s objectives outlined in the quote above, while respecting Zuckerberg’s reticence for private enterprises to serve as arbiters of the truth.

In our model, information is relayed from original sources via sequences of individuals to an eventual “learner,” who wishes to ascertain the state of the world. Our results apply whether the learner is a fully rational Bayesian facing substantial uncertainty, or a naive learner who is simply influenced by the preponderance of messages. The state of the world and corresponding messages are “1” or “0” (e.g., the state can be “climate change is an urgent concern” or “climate change is not an urgent concern”). With noiseless “word-of-mouth” (oral or digital) communication and sufficiently many sources of conditionally independent information, the receiver learns the true state. However, along each transmission chain, the
message may mutate (deliberately or inadvertently), become unintelligible, or be dropped, reducing the information content of the signals that reach the receiver.

Limiting how many times messages can be relayed (network depth) is key to overcoming the noise that builds up as messages are repeatedly relayed. People see fewer messages when depth is capped, but the messages that they do receive are more informative on average, thereby increasing the overall signal-to-noise ratio and enabling at least partial learning. We derive an optimal cap on depth and show that, if depth can be capped, there is no need to otherwise restrict communication.

But what if depth cannot be capped? For instance, some social media platforms do not track whether a message is new or forwarded. Such platforms can still limit network breadth by making it more difficult to forward a message to many people. Such restrictions improve learning, since decreasing breadth increases the relative number of messages from closer sources nodes, as higher breadth increases the expansion properties of a network and thus the relative number of distant to nearby nodes.

Indeed, breadth-limits have been adopted by online messaging platforms. For instance, WhatsApp has capped the number of people that someone can message, for the express purpose of curbing the spread of false information. Facebook has implemented a similar strategy, capping the number of people or groups to which a message can be forwarded.

Some Background Observations  Consider the children’s game of “Telephone,” in which a message is whispered from one player to the next. The final message typically bears little resemblance to the original because of “mutations” along the transmission chain. Such mutations have been found to occur frequently. In Adamic, Lento, Adar and Ng’s (2016) study of online viral memes, one meme was reposted more than 470,000 times, with a mutation rate of around 11 percent and more than 100,000 variants. Indeed, 121 of the 123 most viral memes each had more than 100,000 variants. In particular, when relayed, a message can change – intentionally or inadvertently – in ways that alter its meaning (“mutate”) or render it unintelligible and/or irrelevant to the underlying state (“drop”), or simply not be forwarded (another way of being dropped). Simmons, Adamic and Adar (2011) discuss an interesting example of a message whose meaning was inadvertently lost in transmission. An initial tweet, “Street style shooting in Oxford Circus for ASOS and Diet Coke. Let me know if you’re around!”, was an invitation for people to join the crowd for a commercial being filmed in London. This was misunderstood and within minutes had mutated to “Shooting in progress in Oxford Circus? What?” and then retweeted as

---

1We focus on a simple case in which the platform sets a cap on the number of people to whom a given message is forwarded, similar to WhatsApp’s policy. But our analysis equally applies to any policy that reduces average degree.

2https://about.fb.com/news/2020/09/introducing-a-forwarding-limit-on-messenger/ retrieved Sept 7, 2020.

3Other examples of mutating messages include mythology and the morphing of religious texts. Gurry (2016) estimates that there are around 500,000 textual variants of the Greek New Testament, not including spelling errors.
“Shooting in progress in Oxford Circus, stay safe people.”

Tabloids are rife with examples of intentional message mutations. For example, the Yellowstone Volcano Observatory (YVO) discussed in an article how they repaired a monitoring station that was damaged during a storm. However, aspects of the article were quickly grabbed by a tabloid website, and within hours they had produced their own article that exaggerated and misconstrued the original information and implied Yellowstone was unmonitored. That story even failed to correctly copy and paste the original text. As an example, the word “borehole” in the original was misspelled as “boreal” in the tabloid version. The exact same misspelling was perpetuated in social-media reports and by sources that repeatedly conduct misinformation campaigns about Yellowstone. They either did not know, or ignored, the fact that YVO was the original source of the article, and that the tabloid changed the article’s meaning to something ominous.4

In our model, both intentional and inadvertent changes in content are captured by the probability of nodes changing messages when they forward them, while failures to forward and garbling of content that renders a message uninterpretable are captured by the probability of a message being dropped. Importantly, we allow mutations to occur in different directions with different probabilities. For example, as we have noted, some mutations may be intentional as ideologues relay messages that they prefer telling rather than what they truly heard (e.g., “fake news”), and there may be more ideological pressure in one direction or another depending on the topic.5

In Jackson et al. (2021), we show that in the presence of such mutations, slight uncertainty over the mutation rates severely limits what even a fully Bayesian receiver can learn from a distance, no matter how many independent chains of message relays they hear. Moreover, it appears that people are far from the ideal Bayesian social learner, and tend to be swayed by the preponderance of messages (e.g., see Chandrasekhar et al. (2020)). Indeed, experimental evidence suggests that simply repeating falsehoods makes people more likely to believe them, even when no additional evidence is presented and they have prior knowledge to the contrary (see Fazio et al. (2015) and the references therein). Also, Serra-Garcia and Gneezy (2021) show that people are poor at distinguishing what is true and false, but become more confident in false messages if shared by others. In light of these findings, we step back from particular specifications of sender intentions and receiver updating rules and focus directly on the problem of how platforms can increase message fidelity.

Several governments have increased regulations and fines for disinformation (Posner (2019)). This puts platforms that are hesitant to be “arbiters of the truth” in a difficult situation. Our analysis highlights a policy that such platforms could follow to ensure that most messages on their platforms are true, without needing to police message content. Limiting message passing does not eliminate false information but can improve the overall quality of what gets shared and allow people to learn more from the content they encounter.

---

4https://www.usgs.gov/center-news/playing-telephone-miss-information
5For other perspectives on the role of biased agents in the spread of false information, see Acemoglu, Ozdaglar and ParandehGheibi (2010) and Bloch, Demange and Kranton (2018).
2 A Model of Noisy Information Transmission

Information is relayed by “word-of-mouth” (oral, written, via social media, etc.) from original sources to a “learner.” For instance, the learner may hear from friends about whether there is a link between vaccines and autism and then decide whether to vaccinate their child.

There are two possible states of the world, $\omega \in \{0, 1\}$. Let $\theta$ be the prior probability that the state is 1. Agents are in an infinite network. Each agent is an original source with direct information about the state with probability $r > 0$, and gets no direct information with probability $1 - r$. An original source observes a noisy signal of the state in $\{0, 1\}$, as described below. All original sources then transmit their information to all of their neighbors as a “message”, who then relay it to their neighbors and so forth.

We examine information processing from the perspective of any given node, the learner. We focus on each piece of information traveling along a shortest path to the learner, and abstract from multiple paths and cycles.

We thus model messages as passing along paths in an infinite tree, namely, a random tree (standard Galton-Watson branching process with the learner at the root) with expected degree $k$.

Consider a path from an original source to the learner. Label the source node 1 and the learner $T > 1$, with the message passing through the sequence (or “chain”) of agents $\{1, 2, \ldots, T\}$. Define $m_0 \equiv \omega$ to be the true state.

Messages $m_t$ sent by agent $t$ take on the values $\{0, 1, \emptyset\}$, with the “null message,” $m_t = \emptyset$, indicating that either the message was dropped at some earlier stage or was so garbled that it was not interpretable. In particular, if agent $t \geq 1$ receives a null message $m_t = \emptyset$, then all subsequent agents (including the learner) also receive a null message.

Noisy transmission along the chain proceeds according to the following procedure. If agent $t > 1$ receives a message $m_t \in \{0, 1\}$, then that agent passes along a non-null message ($m_{t+1} \neq \emptyset$) with probability $p$ and the message is dropped/garbled ($m_{t+1} = \emptyset$) with probability $1 - p$.

Each time a non-null message is transmitted, that message mutates from 0 to 1 with probability $\mu_{01} \in [0, 1/2)$, or from 1 to 0 with probability $\mu_{10} \in [0, 1/2)$.

---

6The learner may have information from sources outside of its network. If these sources are not direct, then they can be modeled as part of the network. Otherwise, we can think of this external information as being reflected in the prior.

7We focus on a binary world to crystallize the main ideas. Extensions to richer state spaces and signal structures are left for future research.

8We assume that the initial message $m_1$ is derived from the true state in the same way as any other $m_{t+1}$ depends on $m_t$, as if nature were “agent 0” in the chain. This simplifies expressions, but our analysis easily extends to allow first-signal accuracies and dropping rates to differ from subsequent ones.

9In another paper, Jackson et al. (2021), we analyze social learning when the pass-rate $p$ depends on message content and characterize conditions under which learning can be enhanced by such dependence. However, those results shed little additional light on the role of network structure, and so we focus here on the case in which all messages have the same pass-rate.

10In practice, agents pass along messages for a variety of reasons, be it to inform or persuade or mislead others, tell a joke, or even to just keep a record of what they heard. Since our goal is to understand the type
Thus, for each \( t \geq 0 \): If \( m_t = 1 \), then agent \( t + 1 \) passes along the message \( m_{t+1} = 1 \) with probability \( p(1 - \mu_{10}) \), \( m_{t+1} = 0 \) with probability \( p\mu_{10} \), and \( m_{t+1} = \emptyset \) with probability \( 1 - p \). Similarly, if \( m_t = 0 \), then \( m_{t+1} = 1 \) with probability \( p\mu_{01} \), \( m_{t+1} = 0 \) with probability \( p(1 - \mu_{01}) \), and \( m_{t+1} = \emptyset \) with probability \( 1 - p \). If \( m_t = \emptyset \), then \( m_{t+1} = \emptyset \).

This defines a \( 3 \times 3 \) Markov chain in which \( \emptyset \) is an absorbing state.

(a) A society with a learner and other agents who could pass information to the learner.
(b) Three nodes get initial information of 1 and one gets initial information of 0, while others get no direct information.
(c) Messages pass to the learner. One of the messages was dropped, while three others made it to the learner. One message mutated along its path.

Figure 1: Overall, from four original sources, the learner hears three messages, of which one mutated.

The interesting case is when mutation rates differ across states. If \( \mu_{01} = \mu_{10} \), then messages always are (slightly) more likely to match the starting state and there is no limit to the distance at which messages still allow for some inferences about the state, even at a vanishing rate and even if these mutation rates are unknown. But, for instance, if \( \mu_{01} < \mu_{10} \), then as \( T \) becomes large, any surviving message is more likely to be a 0 than a 1, regardless of the starting state.

In particular, letting \( M = 1 - \mu_{01} - \mu_{10} \), if the state \( \omega = 0 \), the frequency of true (0) to
false (1) messages originating $T$ steps away is easily checked to be

$$\frac{\mu_{10} + \mu_{01} M^T}{\mu_{01} - \mu_{01} M^T}. \tag{1}$$

If $\omega = 1$, the frequency of true (1) to false (0) messages originating $t$ steps away is,

$$\frac{\mu_{01} + \mu_{10} M^T}{\mu_{10} - \mu_{10} M^T}. \tag{2}$$

A Bayesian learner who knows the exact mutation rates ($\mu_{01}, \mu_{10}$) and the exact distance $T$ that a message has traveled can leverage the difference between the frequency ratios, (1) and (2), to glean information about the state from a message, even if it has traveled from a great distance. For such a learner, having more messages from any distance, and hence a broader and deeper network, is unambiguously good for learning. However, perfect knowledge about message distances and mutation rates is unrealistic: people often cannot tell how far away the original source of message is or how likely others are to alter messages. Bearing this in mind, we analyze situations in which the learner cannot tell how far a message has traveled and faces uncertainty about mutation rates.\footnote{In \cite{Jackson2021}, we characterize how agents learn with known mutation rates and known chain lengths. We show that learning from distant messages requires not only sufficient growth in the network, but also that the learner has no uncertainty about the mutation rates. With any uncertainty about mutation rates, it is impossible to learn from any number of messages if they have traveled a sufficiently long distance.}

Finally, we assume that the learner follows the majority of messages received. This can be justified in several ways. For instance, suppose that the learner has a diffuse and symmetric prior on the state and on mutation rates and, after receiving messages, takes a binary action $a \in \{0, 1\}$ that yields payoff +1 if it matches the state or payoff −1 if not. Such an agent would, after Bayesian updating, find it optimal to choose the action corresponding to the most common message received. That said, we do not assume that the learner is Bayesian. For instance, our analysis also applies to a naive learner who uses a rule of thumb of following the majority of messages. The analysis is easily extended to non-symmetric settings where there is some threshold applied that differs from a majority (e.g., the learner has a status quo bias and presumes it is state 0 unless messages of 1 outnumber messages of 0 by some number or by some ratio), but the symmetric case makes the expressions most transparent.

## 3 Improving Learning by Restricting Communication

We analyze how restricting the distance that messages can travel (“depth”) or reducing the network’s degree (“breadth”) improves learning. To gain intuition, consider a situation in which $pk > 1$ so that the learner receives exponentially more messages from farther away. Whenever the mutation rates $\mu_{01}, \mu_{10}$ are unequal, a majority of messages from sufficiently far away are more likely to match the mutation-favored state (state 0 if $\mu_{01} > \mu_{10}$ or state 1
if \( \mu_{01} > \mu_{10} \), causing the learner to take the mutation-favored action regardless of the true state. The learner would be better off in a network with no re-transmission at all: depth capped at \( T = 1 \). But such a draconian restriction might not be optimal, since the learner would receive too few messages.

We examine how to optimally restrict communication with respect to two related objectives. Ultimately, we are most interested in maximizing expected learner welfare, which simplifies to maximizing the probability that the majority of received messages match the true state. Our analysis becomes intractable under that objective, however, so we focus first on the closely-related objective of maximizing the expected number of true minus false messages\(^\text{12}\). We then return to the probability of true messages exceeding false, showing in a numerical example that the main qualitative findings continue to hold.

Say that a set of sources generates a “preponderance of true messages” if the expected number of true messages from these sources exceeds the expected number of false messages, regardless of the state. Whether the network generates a preponderance of true messages depends on the mutation rates (and other model parameters), which may differ depending on message topic, etc. Bearing that in mind, we take a max-min approach, focusing on maximizing worst-case performance given any mutation rates \((\mu_{01}, \mu_{10})\) in an admissible, compact set of mutation rates \( \mathcal{A} \). Say that a set of sources generates a “robust preponderance of true messages” if there are more true than false stories from these sources, on average, given all admissible mutation rates \((\mu_{01}, \mu_{10}) \in \mathcal{A} \) and regardless of the state.

In a platform designed to generate a robust preponderance of true messages, a learner who simply follows the majority of messages will be correct most of the time, given any admissible mutation rates and regardless of the state.

### 3.1 Capping Depth

Proposition\(^\text{11}\) establishes a critical distance \( T^* \) beyond which messages are more likely to be false than true in the mutation-disfavored state.

**Proposition 1** Sources generate a robust preponderance of true messages if and only if they are at distance less than

\[
T^* \equiv \min_{(\mu_{01}, \mu_{10}) \in \mathcal{A}} \frac{\log(\frac{1}{2}) + \log(1 - \min\{\mu_{01}, \mu_{10}\})}{\log(1 - \mu_{01} - \mu_{10})}.
\]

This implies that the expected number of true minus false messages for the worst-case mutation rates (and corresponding state) is maximized when network depth is capped so that the

---

\(^{12}\)Whenever the number of potential sources that are not too far from the learner is reasonably large, maximizing the expected number of true minus false messages also approximately maximizes the probability that true messages outnumber false ones. The reason, effectively, is that increasing how far the mean of the distribution of “excess number of true messages” is above zero also tends to reduce its tail probability of being below zero. Extra complications arise in our context, however, because depth and breadth caps impact the variance of this distribution and these caps are themselves subject to an integer constraint.
learner hears only from sources at distance $T < T^*$. 

The proof of this proposition is relatively straightforward, and so we outline it here. The first part of this proposition follows directly from considering the values of $T$ for which the ratios (1) and (2) are greater than 1 for any admissible pair of mutation rates. To see how the last statement of Proposition 1 follows, consider a platform designer who would like to maximize the expected number of true messages minus the expected number of false messages in the mutation-disfavored state, can block messages from any subset of sources, and knows the mutation rates $(\mu_{01}, \mu_{10})$. Which sources should have their messages blocked and which should be allowed to reach the learner? For every $T < T^*(\mu_{01}, \mu_{10})$ 

$$
T^*(\mu_{01}, \mu_{10}) \equiv \frac{\log\left(\frac{1}{2}\right) + \log\left(1 - \min\{\frac{\mu_{01}}{\mu_{10}}, \frac{\mu_{10}}{\mu_{01}}\}\right)}{\log(1 - \mu_{01} - \mu_{10})},
$$

allowing messages from sources at distance $T$ from the learner increases the expected number of true messages more than the expected number of false ones, regardless of the state. On the other hand, messages from farther away than $T^*(\mu_{01}, \mu_{10})$ are more likely to be false than true in the mutation-disfavored state, thereby increasing the expected number of false stories more than true ones in that state. A platform designer who wished to maximize the expected number of true minus false messages in the worst case would therefore choose to cap depth at $T^*$, allowing messages only from sources closer than $T^*$.

The magnitude of the cap $T^*$ depends on how asymmetric the mutation rates are. On one extreme when $\mu_{01}$ and $\mu_{10}$ are equal, we have $T^* = \infty$. On the other extreme when one of the mutation rates equals $1/2$ and the other equals zero, we have $T^* = 1$.

In principle, the set of possible mutation rates $A$ can be determined by a platform that has data on different types of messages, what we will refer to as “topics.” Setting a depth cap of $T^*$ ensures that the expected number of true minus false messages is positive, on every topic and in every state. That said, if the platform designer can set different depth caps (and knows the mutation rates) for each topic, then setting a customized depth cap of $T^*(\mu_{01}, \mu_{10})$ – with the same formula above – for each topic maximizes the expected number of true minus false messages that users receive, topic by topic.

### 3.2 Capping Breadth

If the depth of the network has been capped at $T \leq T^*$, all received messages are more likely to be true than false. A platform designer intent on maximizing the expected number of true minus false messages would therefore not restrict breadth $d$.

But what if the network’s depth $T$ cannot be capped? Capping $T$ requires following the life-cycle of a message, which can be infeasible in practice. For instance, it may be difficult or impossible to tell whether someone is mutating a previous message she heard or originating a new message. Alternatively, a designer maintaining user privacy may not even observe the messages being forwarded, which are encrypted on some platforms. When depth cannot be restricted, limiting breadth can help ensure that most messages are true. A designer can lower the relative number of long chains by restricting the number of others to whom any given agent can send or forward messages.
Consider a tree with no cap on depth (i.e., an infinite tree) in which each node has average in-degree and out-degree \( k > 0 \). A cap on message forwards of \( d \leq k \) limits each agent to passing along messages to at most \( d \) of its out-neighbors. We suppose that agents follow such a cap by choosing their \( d \) out-neighbors uniformly at random. (The cap is not binding if the agent has \( d \) or fewer friends.)

**Proposition 2** The average degree that maximizes the expected number of true minus false messages for the worst-case mutation rates (and corresponding state) is

\[
d^* = \min_{(\mu_{01}, \mu_{10}) \in A} \frac{1 - Z}{p (1 - (1 - \mu_{01} - \mu_{10}) Z)},
\]

where \( Z = \left( \frac{\max(\mu_{10}, \mu_{01}) - \min(\mu_{10}, \mu_{01})}{2 \max(\mu_{10}, \mu_{01})(1 - \mu_{01} - \mu_{10})} \right)^{1/2} \).

The term \( pd \) captures the expansion property of a network with average degree \( d \), with far-away messages dominating what the learner receives if \( pd > 1 \). Not surprisingly, the optimal breadth cap \( d^* \) is always low enough so that \( pd^* \leq 1 \), with \( pd^* = 1 \) only in the special case with symmetric mutation rates, \( \mu_{01} = \mu_{10} \). (To see why, note that \( Z = 0 \) if and only if \( \mu_{01} = \mu_{10} \) and \( 1 - \mu_{01} - \mu_{10} \in (0, 1) \).)

As the likelihood \( p \) that messages are passed is decreased, fewer messages come from longer distances and the optimal cap is higher, meaning that people are less constrained in how widely they can forward any given message. In the extreme as \( p \) approaches 1, so messages are likely to travel far, the optimal cap is set such that the expected degree is less than 1.

If we increase the higher of the two mutation rates, then \( Z \) increases and \( (1 - \mu_{01} - \mu_{10}) \) decreases. This implies that \( \frac{1 - Z}{1 - (1 - \mu_{01} - \mu_{10}) Z} \) decreases, since the denominator shrinks more slowly, and a tighter cap is needed. In contrast, when the lower of the two mutation rates increases (but not so much to pass the other), a looser cap is possible. In other words, caps are less useful when mutation rates come more into balance.

Proposition 1 showed that when restricting depth, the cap that generates a robust preponderance of messages coincides with the cap that maximizes the extent to which truthful messages outnumber false ones on average. But this is not the case when restricting breadth: the cap given in Proposition 2 only achieves the latter goal. A looser breadth cap is possible if the goal is merely to ensure that false messages never outnumber true messages in expectation, in the sense that the network as a whole has a robust preponderance of true messages. Proposition 3 characterizes the maximal breadth \( \bar{d} \) for an infinite-depth tree to have this property.

\[13\] In the special case with symmetric mutation rates and no dropping (\( p = 1 \)), the optimal breadth limit \( d^* = 1 \), meaning that the optimally-pruned tree is a single infinite chain. Otherwise whenever \( \mu_{01} \neq \mu_{10} \), the optimal breadth-limit prunes the tree sufficiently that the expected number of sources is always finite.
Proposition 3 An infinite-depth network generates a robust preponderance of true messages if and only if its degree is less than
\[
\bar{d} = \min_{(\mu_{01}, \mu_{10}) \in A} \frac{1 - 2 \max\{\mu_{01}, \mu_{10}\}}{p(1 - \mu_{01} - \mu_{10})}.
\]

As can be easily checked, the breadth cap \( \bar{d} \) always exceeds the cap \( d^* \). This difference highlights a potential conflict of interest between platforms and their users. In particular, if a platform seeks to maximize its overall volume of communication, subject to the constraint that there is a robust preponderance of the truth, then the platform will set a looser cap on network breadth than a social planner seeking to maximize learner welfare.

3.3 Probabilities that True Outnumber False Messages

The above results examine the expected number of true minus false messages, but more relevant for the learner is the probability that true messages exceed false ones. This is not tractable in closed form, but we observe through numerical simulations that the comparative statics of varying network depth/breadth are similar to those of the caps above. In particular, the probabilities of having true messages exceed false ones for different combinations of \( d \) and \( T \) (for \( d \)-regular trees) are pictured in Figure 2.

![Graphs showing probabilities of true messages exceeding false messages](a) Probability that more messages are true than false in state 0, as a function of the cap on tree depth for three values of breadth. (b) Probability that more messages are true than false in state 0 for three relative mutation rates (all with depth 10).

Figure 2: These graphs show how the probability that most received messages are true in state 0 (the state more likely to have mutations), varies with the depth and breadth of the learner’s social network. The parameters used in these simulations are \( p = .2, \mu_{01} = .10 \) and \( \mu_{10} = .03 \) for panel (a) and \( p = .2, \mu_{01} = .10 \) and \( \mu_{10} \in \{.03, .05, .07\} \) for panel (b).

In the numerical examples illustrated in Figure 2, messages of any distance are more likely to be true than false in state 1, the mutation-favored state. However, once a message travels too far, it is more likely to be false than true in state 0. By Proposition 1, the threshold depth at which false messages on average overtake true ones is \( T^* = 7.5 \), regardless of
network breadth. The breadth-cap in Proposition 2 that maximizes the extent to which true messages outnumber false ones is \(d^* = 1.83\), while the cap in Proposition 3 that ensures a robust preponderance of the truth is \(d = 4.6\).

Those thresholds do not maximize the probability that true messages outnumber false ones. Nonetheless, as is clear from Figure 2, these are also close to the maximizers of probability.

Figure 2 also shows that as the mutation rates become closer to each other, then it could be optimal to avoid capping breadth (depth), so the confound in learning originates from the asymmetry in mutation rates, which then increasingly biases information the longer it travels.

### 3.4 Homophily

A common feature of many social networks is that nodes tend to be more similar to neighbors than they are to those more distant to them. In communication networks, nearby nodes may share similar tendencies to mutate or drop messages. For example, one cluster of users may be primarily responsible for exaggerating information in one direction, while another group may be responsible for another.

Given our focus on robust bounds for a set of admissible mutation rates, the caps we derive can account for this sort of heterogeneity. Nevertheless, there are interesting questions for further study. Optimizing learning within homophilous networks involves more than just trading off signal and noise. For instance, if clusters of nodes in different parts of the network tend to mutate messages in different directions, then loosening caps on depth and/or breadth could help receivers hear messages from other groups. Learning could then be non-monotone in the caps.

### 3.5 Discussion

Our results above show platforms can reduce the frequency of false communications by limiting either the depth or breadth of people’s information networks. These policies enable agents to partially learn the true state in the presence of asymmetric mutations, by ensuring that relatively more of an agent’s received messages originate from nearby.

Most importantly, such policies improve learning without requiring message content to be observed, an attractive feature for platforms intent on respecting user privacy. Even if messages are observable, a platform designer who does not know the true state of the world may not be able to discern true messages from false ones, or may prefer to take a content-neutral stance and avoid censoring messages.

Cutting down on “fake news” has recently become an expressed objective for social media platforms that have been blamed for facilitating the spread of misinformation. For example,

---

14 The breadth figures are all done for a fixed depth of \(T = 10\) rather than an infinite depth, for the sake of tractable simulations.
the messaging platform WhatsApp was criticized for allowing the spread of incendiary rumors leading to mob killings in India and for serving as a vehicle of misinformation in 2018 Brazilian elections. Similarly, WhatsApp placed a cap of 5 on the number of people to whom a message can be forwarded in an attempt to curtail the spread of false information \cite{Hern and Safi (2019)}.

4 Concluding Remarks

We introduced a model of social learning via relayed messages in the presence of noise. Even slight uncertainty about how messages mutate leads to a tradeoff between the size of someone’s network and their ability to learn. In that context, capping the depth of a network – or the breadth if the depth cannot be capped – improves learning; and without the need for censoring based on message content. Nonetheless, as our results show, a profit-motivated platform might still prefer different caps than would be socially optimal.

The challenges of learning from word-of-mouth communication can also motivate learners to seek out information from closer, trusted contacts. To the extent that a platform can make it easy to trace the length of a chain and a message’s history, that could also enhance learning. However, tracing information backward is difficult in many cases, especially for platforms with encrypted messages. In such cases, caps on the breadth of users’ word-of-mouth networks can be powerful tools, and relatively easily implemented.

Caps reduce the relative prevalence of false messages because messages originating closer to the receiver have fewer chances to become distorted. Of course, some primary sources may also be originating false information, perhaps intentionally for ideological reasons. Caps reduce the scope for such messages to go viral. To the extent that this diminishes ideological sources’ incentive to originate false content more than it dissuades reliable sources from originating true content, the positive impact of caps is understated in our model, which treats information arrival as exogenous. Fully analyzing sources’ incentives when originating information is a topic for further study.\footnote{If sources that wish to deliberately manipulate beliefs have choices of where to appear in the network, or simply happen to appear more frequently in some parts of the network, that could introduce asymmetries in exposure to true versus false information across learners, which could result in more complex policies, which is also an interesting topic for further study.}

More broadly, this paper poses the question of how to design policies that curb misinformation while sidestepping the issue of determining what is true or false. Whether out of reluctance to taking a stance, out of respect for privacy in communication, or simply an inability to determine truth value of content at scale, there seems to be a demand for such information agnostic policies. We identify a range of information agnostic policies that improve learning in the context of noisy communication, whereby either the depth or breadth of people’s communication networks are restricted. It is left to future research to consider whether other information agnostic policies, such as revealing a message’s origin and trajectory, can further curb misinformation.

\footnote{Relatedly, if sources that wish to deliberately manipulate beliefs have choices of where to appear in the network, or simply happen to appear more frequently in some parts of the network, that could introduce asymmetries in exposure to true versus false information across learners, which could result in more complex policies, which is also an interesting topic for further study.}
5 Proofs

The proof of Proposition 1 is straightforward and outlined in the main text. Here we provide proofs of Propositions 2 and 3. We presage those proofs with a lemma that provides some expressions that are useful Markov Chain formulas.

**Lemma 1** Suppose that \( p > 0 \) and consider any mutation rate \( \mu_{01}, \mu_{10} \in (0, 1/2) \). If the state is 0 and agent \( t \geq 1 \) receives a non-null message, then the message is 0 (matching the true state) with probability

\[
X_0(t) = \frac{\mu_{10} + \mu_{01} M^t}{\mu_{10} + \mu_{01}}.
\]

If the state is 1 and agent \( t \geq 1 \) receives a non-null message, the message is 1 (matching the true state) with probability

\[
X_1(t) = \frac{\mu_{01} + \mu_{10} M^t}{\mu_{01} + \mu_{10}}.
\]

It follows that

\[
X_0(t) - (1 - X_1(t)) = M^t.
\]

As \( t \) grows, regardless of the starting state, the probability that a surviving message is a 0 and a 1, respectively are

\[
\pi_0 = \frac{\mu_{10}}{\mu_{10} + \mu_{01}} \quad \text{and} \quad \pi_1 = \frac{\mu_{01}}{\mu_{10} + \mu_{01}}.
\]

Finally, if \( \mu_{01} = \mu_{10} = \mu \), then the message matches the true state with probability

\[
X(t) = \frac{1 + M^t}{2}. \tag{4}
\]

**Proof of Lemma 1** We derive the expressions of \( X_0, X_1 \), which can also be deduced from standard Markov chain results, but it may be useful for the reader to see the derivation. The proof is by induction. We give the proof for \( X_0 \), when the state is 0. The proof for \( X_1 \) is symmetric and the expression for \( X \) is a special case.

First, note that if \( t = 1 \) then this expression simplifies to \( 1 - \mu_{01} \), which is exactly the probability that the message has not mutated, and so this holds for \( t = 1 \).

Then for the induction step, supposing that the claimed expression is correct for \( t - 1 \), we show it is correct for \( t \).

The probability of matching the true state at \( t \) is the probability of not matching at \( t - 1 \) times \( \mu_{10} \) plus the probability of matching at \( t - 1 \) times \( 1 - \mu_{01} \), which by the induction assumption can be written as

\[
\left[1 - \frac{\mu_{10} + \mu_{01} M^{t-1}}{\mu_{10} + \mu_{01}}\right] \mu_{10} + \left[\frac{\mu_{10} + \mu_{01} M^{t-1}}{\mu_{10} + \mu_{01}}\right] (1 - \mu_{01})
\]
\[
\begin{align*}
&= \mu_{10} + \left[ -\frac{\mu_{10}^2 - \mu_{01}\mu_{10}M^{t-1} + \mu_{10} - \mu_{10}\mu_{01} + \mu_{01}M^{t-1} - \mu_{10}^2M^{t-1}}{\mu_{10} + \mu_{01}} \right] \\
&= \mu_{10} + \left[ -\frac{\mu_{10}^2 + \mu_{10} - \mu_{10}\mu_{01} + \mu_{01}M^{t-1}(1 - \mu_{10} - \mu_{01})}{\mu_{10} + \mu_{01}} \right] \\
&= \frac{\mu_{10} + \mu_{01}M^{t}}{\mu_{10} + \mu_{01}}.
\end{align*}
\]

as claimed. \(\blacksquare\)

**Proof of Proposition 2:**

Suppose without loss of generality that \(\mu_{01} \leq \mu_{10}\). Given this ordering, the worst case for the number of true minus false messages occurs when the true state is 1.

When the average degree of the Galton-Watson tree is \(d\), it follows by induction and iterated expectations that there are \(d^t\) nodes at a distance \(t\) away from the learner, in expectation. Therefore, the expected number of true messages received is \(\sum_{t=1}^{\infty} r(pd)^t X_1(t)\), and the expected number of false messages received is \(\sum_{t=1}^{\infty} r(pd)^t(1 - X_1(t))\). The expected number of true minus false messages is

\[
\sum_{t=1}^{\infty} r(pd)^t (X_1(t) - (1 - X_1(t))).
\]

Thus we need to maximize \(\sum_{t=1}^{\infty} (pd)^t(2X_1(t) - 1)\); i.e.,

\[
\sum_{t=1}^{\infty} (pd)^t\left(\frac{2\mu_{10}M^t + \mu_{01} - \mu_{10}}{\mu_{10} + \mu_{01}}\right).
\]

The first order conditions are

\[
\sum_{t=1}^{\infty} tp^t d^{t-1} \left(\frac{2\mu_{10}M^t + \mu_{01} - \mu_{10}}{\mu_{10} + \mu_{01}}\right) = 0. \tag{5}
\]

Multiplying this by \(d\), this is equivalent to

\[
\sum_{t=1}^{\infty} tp^t d^t \left(\frac{2\mu_{10}M^t + \mu_{01} - \mu_{10}}{\mu_{10} + \mu_{01}}\right) = 0
\]

or

\[
\sum_{t=1}^{\infty} tp^t d^t M^t \left(\frac{2\mu_{10}}{\mu_{10} + \mu_{01}}\right) - \sum_{t=1}^{\infty} tp^t d^t \left(\frac{\mu_{10} - \mu_{01}}{\mu_{10} + \mu_{01}}\right) = 0.
\]

Noting that \(\sum_{t=1}^{\infty} tz^t = z/(1 - z)^2\), we rewrite the above as

\[
\frac{pdM^2\mu_{10}}{(1 - pdM)^2} - \frac{pd(\mu_{10} - \mu_{01})}{(1 - pd)^2} = 0.
\]
Rearranging terms leads to the claimed expression. To check the second order conditions, note that the second derivative is

\[
\sum_{t=2}^{\infty} (t-1)tp'\,d^{t-2} \left( \frac{2\mu_{10} M^t + \mu_{01} - \mu_{10}}{\mu_{10} + \mu_{01}} \right).
\]

Multiplying this by \(d\) does not change the sign, and gives

\[
\sum_{t=1}^{\infty} (t-1)tp'\,d^{t-1} \left( \frac{2\mu_{10} M^t + \mu_{01} - \mu_{10}}{\mu_{10} + \mu_{01}} \right),
\]

which is the same as (5), but where the weights are adjusted by \((t-1)\). Since the expression \((2\mu_{10} M^t + \mu_{01} - \mu_{10})\) is decreasing in \(t\) (and eventually negative), then by reweighting with weights \((t-1)tp'\,d^{t-1}\) compared to \(tp'\,d^{t-1}\) puts more weight on higher terms (in a dominance sense), and since (5) is 0, the second derivative expression is necessarily negative.

**Proof of Proposition 3:**

Suppose without loss of generality that \(\mu_{01} \leq \mu_{10}\).

By Lemma 1, the expected number of true messages exceed the expected number of false messages when the state is 0, since for every \(t\), \(X_0(t) = \frac{\mu_{10} + \mu_{01} M^t}{\mu_{10} + \mu_{01}} > \frac{\mu_{01} - \mu_{01} M^t}{\mu_{10} + \mu_{01}} = 1 - X_0(t)\).

It remains to be shown that the expected number of true messages also exceed the expected number of false messages when the state is 1. Using the expressions from the proof of Proposition 2, we need to characterize the conditions under which

\[
\sum_{t=1}^{\infty} r(pd)^t X_1(t) > \sum_{t=1}^{\infty} r(pd)^t (1 - X_1(t)).
\]

Thus we need \(\sum_{t=1}^{\infty} (pd)^t (2X_1(t) - 1) > 0\); i.e., when

\[
\sum_{t=1}^{\infty} (pd)^t \left( \frac{2\mu_{10} M^t + \mu_{01} - \mu_{10}}{\mu_{10} + \mu_{01}} \right) > 0.
\]

If \(\mu_{01} = \mu_{10}\), then this holds regardless of \(d\), which establishes the first statement of the proposition. Next, suppose \(\mu_{01} < \mu_{10}\). Then it follows that

\[
\sum_{t=1}^{\infty} (pd)^t \left( \frac{2\mu_{10} M^t + \mu_{01} - \mu_{10}}{\mu_{10} + \mu_{01}} \right) > 0
\]

\[
\iff \sum_{t=1}^{\infty} (pMd)^t (2\mu_{10}) > \sum_{t=1}^{\infty} (pd)^t (\mu_{10} - \mu_{01})
\]

\[
\iff \frac{pMd}{1 - pMd} (2\mu_{10}) > \frac{pd}{1 - pd} (\mu_{10} - \mu_{01})
\]

\[
\iff \frac{2\mu_{10} M}{1 - pMd} > \frac{\mu_{10} - \mu_{01}}{1 - pd}
\]
\[
\iff \frac{1 - pM d}{2\mu_{10} M} < \frac{1 - pd}{\mu_{10} - \mu_{01}} \\
\iff \frac{1}{2\mu_{10} M} - \frac{pd}{2\mu_{10}} < \frac{1}{\mu_{10} - \mu_{01}} - \frac{pd}{\mu_{10} - \mu_{01}} \\
\iff \frac{1}{p} \left( \frac{1}{2\mu_{10} M} - \frac{1}{\mu_{10} - \mu_{01}} \right) < d \left( \frac{1}{2\mu_{10}} - \frac{1}{\mu_{10} - \mu_{01}} \right) \\
\iff d < \frac{1}{p} \frac{\mu_{10} - \mu_{01}}{\mu_{10} - \mu_{01}} - 2\mu_{10} \\
\iff d < \frac{1}{p} \frac{2\mu_{10} + \mu_{01} - \mu_{10}}{\mu_{10} + \mu_{01}} \\
\iff d < \frac{1}{pM} \frac{\mu_{10} + \mu_{01} - 2\mu_{10}(\mu_{10} + \mu_{01})}{\mu_{10} + \mu_{01}} \\
\iff d < \frac{1 - 2\mu_{10}}{pM}.
\]

This is the condition in the proposition for \(\mu_{01} \leq \mu_{10}\), which was without loss of generality, concluding the proof. ☐
References

Acemoglu, Daron, Asuman Ozdaglar, and Ali ParandehGheibi, “Spread of (mis) information in social networks,” *Games and Economic Behavior*, 2010, 70 (2), 194–227.

Adamic, Lada A., Thomas M. Lento, Eytan Adar, and Pauline C. Ng, “Information evolution in social networks,” in “Proceedings of the ninth ACM international conference on web search and data mining” ACM 2016, pp. 473–482.

Bloch, Francis, Gabrielle Demange, and Rachel Kranton, “Rumors and social networks,” *International Economic Review*, 2018, 59 (2), 421–448.

Chandrasekhar, Arun G, Horacio Larreguy, and Juan Pablo Xandri, “Testing models of social learning on networks: Evidence from two experiments,” *Econometrica*, 2020, 88 (1), 1–32.

Fazio, Lisa K, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh, “Knowledge does not protect against illusory truth.,” *Journal of Experimental Psychology: General*, 2015, 144 (5), 993.

Gurry, Peter J, “The Number of Variants in the Greek New Testament: A Proposed Estimate,” *New Testament Studies*, 2016, 62 (1), 97–121.

Hern, Alex and Michael Safi, “WhatsApp puts limit on message forwarding to fight fake news,” Jan 2019.

Jackson, Matthew O., Suraj Malladi, and David McAdams, “The Fragility of Social Learning in the Presence of Noisy Messages,” *mimeo: https://web.stanford.edu/~jacksonnm/mutation.pdf*, 2021.

Posner, Michael, “How Social Media Companies Need To Address Disinformation Globally,” Jun 2019.

Serra-Garcia, Marta and Uri Gneezy, “Mistakes, Overconfidence, and the Effect of Sharing on Detecting Lies,” *American Economic Review*, 2021, 111 (10), 3160–83.

Simmons, Matthew P., Lada A. Adamic, and Eytan Adar, “Memes Online: Extracted, Subtracted, Injected, and Recollected,” *icwsm (International Conference on Web and Social Media Proceedings)*, 2011, 11, 17–21.

Vosoughi, Soroush, Deb Roy, and Sinan Aral, “The spread of true and false news online,” *Science*, 2018, 359 (6380), 1146–1151.