The burden of proof: fact-checking and credibility in communication networks

Robin Lenoir *
École Normale Supérieure de Lyon

November 27, 2020

Abstract

Can fact-checking improve communication in a network? I study a communication network where agents can publicly commit ex-ante to fact-check any message they send with a reliability of their choice. I show that truth-seeking agents use fact-checking as a device to verify information while biased agents -who want false opinions to spread - use fact-checking as a persuasion device to improve their credibility. I describe how a designer can implement full communication (all messages are trusted and transmitted) by choosing an appropriate cost for fact-checking commitment, to be paid by agents. Finally, I study who carries the "burden of proof"; that which agents bear the necessary fact-checking to ensure sufficient trust in the network. I show that when the cost of fact-checking is low, unbiased agents carry the "burden of proof." Conversely, when the cost is high, biased agents carry it.

JEL CLASSIFICATION CODES: D83, D85, C72.

KEYWORDS: Networks, strategic communication, bayesian persuasion, fact-checking, misinformation, social media.

*Email: lenoir.robin@gmail.com. I am deeply thankful to Christophe Bravard and Ben Golub for extensive comments on this project. I also wish to thank Shengwu Li and Jerry Green for their help on early versions of this paper.

1 Introduction

Concerns about misinformation on social media have grown in the public debate in recent years. While many worry about the spread of false information, misinformation may also be responsible for useful information failing to reach the population. In a recent survey about COVID-19 information on social-media, Sarah Kreps and Doug Kriner observed that less than 50% of respondents were able to identify that a correct piece of information was valid [Kreps & Kriner (2020)]. In comparison, more than 70% could identify a false piece of information as incorrect. Users assess the veracity of information acquired on social media based on the credibility of both the information and its source. On social media, this problem is magnified because information circulates among a network of agents, and the primary source is rarely known. As a result, agents need to assess the credibility of both a particular sender and the whole chain of information upstream.

My paper proposes a new tool for platform designers to correct communication failures that arise when agents do not trust each other. More precisely, I build a theoretical model inspired by Bloch, Demange & Kranton (2018), where agents in a network can strategically choose to share a piece of information with an unknown source. I then show how platforms can restore trust and solve communication failures by allowing agents to fact-check the information they send. Until now, most fact-checking initiatives by digital platforms have taken the form of warnings and advice provided to users who *consume* the information rather than produce or relay it. I propose a different approach, where agents who send or transmit information choose to subject *themselves* to fact-checking. Building on insight from the Bayesian persuasion literature, I show that even "biased" agents (who want falsehoods to spread) may be willing to subject themselves to fact-checking to gain credibility. I illustrate how platform designers can incentivize enough fact-checking to fully restore communication within the network.

Following Bloch et al. (2018) model, my paper features two types of agents: biased and unbiased. The existence of these types is common knowledge. **Unbiased** agents want others to hold correct beliefs about some binary state of the world, coded as 1 or 0. In contrast, **biased** agents want to disseminate a specific, potentially false belief: they want others to systematically believe that the state of the world is 1, regardless of its true value. A single agent chosen at random (the "source") learns the state and can emit a signal of their choice (that may be false) to inform their neighbors. If designated as coming from the source, unbiased agents will create signals matching this information about the world's state. Biased agents, on the other hand, will always create a signal equal to 1. Agents who receive the message can then transmit it to other agents but cannot alter it. They can, however, block the message if they deem it untrustworthy. Importantly, agents do not know whether the person who sends them the message is the source or merely transmitting the message. To evaluate the veracity of a message, agents have to assess how likely it is that the source is biased based on the proportion of biased agents in each part of the network. The original contribution of my paper is introducing a fact-checking commitment into this setting. Before communication, all agents are allowed to buy a fact-checking device of **reliability** r, at a price $c \times r$. An agent who buys this device commits to having any message they send fact-checked with reliability r, meaning that a false message has a probability r of being detected as false. If a message is detected as false, then it is blocked by the device.

Such a stylized model of a fact-checking commitment is too simplistic to resemble any particular device a platform could provide. Its tractability, however, allows me to highlight the persuasion forces that more realistic fact-checking commitment tools might generate. In the analysis, I show that both types of agents will use this device, but for different reasons. Unbiased agents will use fact-checking in a traditional way as a means of verifying the information. Biased agents, on the other hand, will use it as a device of **persuasion**. They use fact-checking to improve their credibility and ensure that their message is transmitted. In addition to fact-checking devices on a platform, several interpretations of fact-checking commitment could be made, such as reputation or third-party involvement.

I examine how certain fact-checking strategy profiles $\boldsymbol{r}$ can restore trust and enable communication, depending on the network structure. $\boldsymbol{r}$ corresponds to the vector of "reliabilities" each agent choose for their own fact-checking device. In the baseline, no agents choose to be fact-checked (all reliabilities r are set to 0), and communication failures arise in most network structures. In such a situation, agents assess the credibility of messages solely based on the number of biased agents upstream of them. If the proportion of biased agents is too high, agents perceive messages as untrustworthy and choose to block them.

If agents choose to be fact-checked (with positive reliability), they gain credibility and can persuade skeptical agents to transmit more messages. The gains from credibility are not limited to direct relations. Indeed, agents evaluate a message based on its complete inferred history of transmission. They update their beliefs taking into account not only how much fact-checking the sender of the message is subjected to but also who might have sent the message to the sender who sent the message to them in the first place, and so forth. For all network structures, I determine the existence of a set of **trust-inducing** fact-checking strategies that lead to a state of "**full communication**" in the network. Full communication means that agents believe any messages they receive. There is sufficient trust in the network such that any agent transmits any message.

Finally, a platform designer can incentivize communication using fact-checking commitments. I characterize a range of fact-checking costs for which trust-inducing, fact-checking strategies are chosen by all agents at equilibrium. The expression of this range depends on the network structure. In choosing a precise cost for fact-checking, a platform designer can shift the *burden of proof*, i.e. who performs the necessary fact-checking to ensure trust. For fact-checking at a small cost, it is mainly unbiased agents who fact-check, and they use fact-checking to verify the information. For fact-checking at a high cost, the burden of proof shifts towards biased agents, who use it to persuade other agents of their credibility.

This paper contributes to the literature on strategic communication within networks. The research on strategic diffusion, initiated by "cheap talk" models in Crawford & Sobel (1982) and Green & Stokey (2007), studies games with a Sender who has information and a Receiver whose optimal action depends on this information. Several recent papers have studied how strategic communication is affected by network structures when several agents are involved. Early papers studied how networks might form endogenously from strategic communication ([Calvó-Armengol, Martí & Prat (2015)], [Hagenbach & Koessler (2010)]). Other models study how (exogeneous) restrictions on the communication structure influence cheap-talk outcomes. Galeotti, Ghiglino & Squintani (2013), for example, propose an extension to the "cheap-talk" setting where each agent can communicate with a restricted set of neighbors.

This study most closely relates to a set of papers that characterize how information diffuses over several rounds of strategic communication. Anderlini, Gerardi & Lagunoff (2012) study a repeated game where agents select an action at each period and send a message to all agents to influence their actions in future periods. Ambrus, Azevedo & Kamada (2013), on the other hand, study "hierarchical cheap-talk," where two agents communicate through a chain of strategic intermediators. Each intermediator has an idiosyncratic bias and wants to influence the final action. One particularity of both these models, shared by this paper, is that agents are both senders and receivers of messages. This approach follows the advice of sociologists Paul Lazarsfeld and Elihu Katz in 1966 [Katz & Lazarsfeld (1966)] and considers the individual agent in their "two-fold capacity as a communicator and as a relay point in the network of mass communication"[1]. Recent papers have introduced more general network structures to study information diffusion with "two-fold" communication (agents both receiving and transmitting information strategically). In a working paper, Germán Gieczewski [Gieczewski (2020)] studies a general framework where privately informed agents with heterogeneous biases can emit verifiable signals that are then passed on from agent to agent in a network. My paper departs from this modeling in that the primary source of information (the original sender of the signal) cannot be identified.

As mentioned earlier in the Introduction, the underlying communication process of my model is directly borrowed from Bloch et al. (2018) . I innovate from this paper in two ways. First, a fact-checking commitment gives agents some leverage over their credibility. In Bloch et al. (2018), agents' credibility solely depended on the distribution of types among the network, which was exogenous. My paper endogenizes credibility as a strategic feature of the model. Second, my paper frames the problem of communication failures as a design problem for platform managers. Without fact-checking, Bloch et al. (2018) showed that lack of trust in a network, due to uncertainty regarding the source, can lead to communication failures. I show how platform designers can use fact-checking to fully restore communication.

My paper builds on the work focusing on cheap-talk communication over networks by introducing concepts from the information design literature into these frameworks. By allowing commitment to fact-checking, I introduce the concept of *bayesian persuasion* [Kamenica & Gentzkow (2011)] in a network setting. The idea of Bayesian persuasion is that untrustworthy *senders* can gain credibility by committing to the type of message they send. I show that fact-checking leads to a persuasion situation where agents use it to gain credibility as they transmit messages. The question of how Bayesian persuasion extends to network settings remains largely open. Egorov & Sonin (2019) studies a model where a network of receivers can choose to either buy a biased signal from a central sender (a public institution or a journal) or rely on their network to get the information for free. Candogan (2019) proposes a setting where agents choose an action in a network of strategic complementarities. A centralized designer can send a public message to incentivize action. My setting is original in that the information designers carrying out persuasion are *agents themselves* rather than a centralized entity. In my model, agents have access to a fact-checking commitment they

[1]In their Introduction to *Personal Influence*, Paul Lazarsfeld and Elihu Katz wrote "We have once again become interested in person-to-person communication and it now has become increasingly clear that the person who reads something and talks about it with other people cannot be taken simply as a simile for social entities like newspapers or magazines. He himself needs to be studied in his two-fold capacity as a communicator and as a relay point in the network of mass communication."

can use to shape their credibility. Such endogenous choices shape communication outcomes in equilibrium.

In the next section, I present the baseline model of fact-checking in a communication network. In Section 3, I characterize the set of fact-checking strategies that allow Full Communication within the network. In Section 4, I show how a designer can generate such strategies to achieve Full Communication. Section 5 discusses the results.

2 Model

2.1 Incentives: the voting game

There is a population of $|\mathcal{N}| = n$ agents and two possible states of the world $\theta \in \{0, 1\}$. Agents benefit from a collective decision $x \in \{0, 1\}$. There are two types of agents, distinguished by their preferences over x. **Unbiased agents** wish for x to match the state of the world θ. Their payoffs from x are captured by the quadratic disutility $-\lambda_U(x - \theta)^2$, where $\lambda_U \in \mathbb{R}^+$ scales their **motivation**. **Biased agents** wish that x is set to 1, no matter what θ is. Their payoffs from x are captured by $-\lambda_B(x - 1)^2$.

A uniform probabilistic vote procedure implements the collective decision. Lets z denotes the number of agents who voted for 1. The probability that the outcome of the collective decision is $x = 1$ is equal to $\frac{z}{n}$. Biased agents wish that for a high number of votes for $x = 1$, unbiased agents wish for a high number of votes to match the state.

With such a procedure, agents vote according to their beliefs (see Lemma 1 in [Bloch et al. (2018)]). If unbiased agent i thinks that $\theta = 1$ is more likely, they will vote for 1. It follows directly that biased agents wish that others believe that $\theta = 1$. Unbiased agents, on the other hand, wish that other agents share the same beliefs as they do. Potential interpretation therefore go beyond strict collective decisions with votes. Any setting where agents care that others share their beliefs fits this model. One can use it to model public debate about ethical, political, or cultural issues where agents disagree about an uncertain underlying situation.

2.2 Pre-play communication

To inform their vote, agents communicate amongst a network. A **network** is a pair $G =< \mathcal{N}, \mathcal{E} >$, where $\mathcal{N}$ is the set of agent, and $\mathcal{E}$ the set of directed edges. I write $\mathcal{N} = \mathcal{U} \bigcup \mathcal{B}$ with $\mathcal{U}$ and $\mathcal{B}$ being the set of unbiased and biased agents, respectively. $\mathcal{E}_\mathcal{U}$ and $\mathcal{E}_\mathcal{B}$ denotes the set of outgoing edges where the starting point is an unbiased agent and a biased agent respectively.

I denote ij the directed link from agent i to j. The set N_i is the set of neighbors of i, *i.e.* any agent i' for which ii' is in $\mathcal{E}$. Importantly, I assume that the network is a **tree** throughout the paper: for any i and j, there exists a unique path between i and j. This assumption is made for tractability and plays an important role in solving the model.

2.2.1 Message creation and transmission

Before the vote, agents communicate about the state of the world. Nature emits a private signal with probability p to a single agent chosen uniformly at random. Suppose Nature chooses j. Agent j, referred to as the **source**, is then given the right to send any message $m_{ji} \in \{0, 1\}$ to any of their neighbor i. Throughout the paper, I assume that j (and any other agent) always sends the same message to all of their neighbors simultaneously[2]. We can, therefore, simply denote the outgoing message as m_j. Agent j can send a message matching the signal they receive ($m_j = \theta$) or send a false message ($m_j \neq \theta$). They can also choose not to send any message.

Suppose agent i receives a message m_j from j. Agent i can then pass them on to their neighbors by sending m_i. Agent i is not the source of the message but a **messenger**. Messengers are allowed to transmit any messages they received (they can also choose to block them). They cannot, however, alter it and create original messages as the source does. Formally, if i is a messenger receiving a message from j, we must have $m_i = m_j$ or $m_i = \emptyset$. If i sent the message to his or her neighbors, these neighbors can transmit the message themselves, and so on until no further communication is possible.

2.2.2 Fact-checking commitment

Before any communication (before the source is designated), agents can commit to a **fact-checking device**. This device verifies the trustfulness of messages they send along their outgoing edges.

Consider any agent j (not necessarily the source from before). For each of their outgoing edges ji, j can commit publicly to a fact-checking device of reliability r_{ji}. For this reliability, j will have to pay a cost $c \times r_{ji}$, with $c \geq 0$. Suppose now that the game has started and j receives a message. They decide to transmit this message to their neighbors and sends m_{ji} to i. The fact-checking device verifies this message with reliability r_{ji} before it reaches i. Message m_{ji} can *pass* the test (and be transmitted) or *fail* (and be muted). When the state of the world is $\theta = 0$, a message $m_{ji} = 1$ passes the test with probability $1 - r_{ji}$. A message $m_{ji} = 0$ passes the test with probability equal to 1. Conversely, when the state of the world is $\theta = 1$, a message $m_{ji} = 1$ passes the test with probability 1 and a message $m_{ji} = 0$ passes the test with probability $1 - r_{ji}$. This specification implies that the fact-checking test gives false negatives with probability $1 - r_{ji}$ and never gives false positives. At some steps of the analysis, it will be important to distinguish messages before fact-checking and after fact-checking. I denote m_{ji}^{out} the message sent by j to i before it is fact-checked and m_{ji}^{in} the message received by i from j after it is fact-checked. Both messages have the same value, but it is possible that m_{ji}^{in} does not exist. If there is no ambiguity, I will drop the superscript.

If the transmission is successful, i receives the message and can pass it on to their neighbors. It is important to note that j committed before the beginning of the game. It means that if j wants to transmit a false message (because they are a biased source, for example,

[2]This assumption is not restrictive. Without the assumption, a similar behavior would arise endogenously from strategic incentives. Intuitively, if an agent passes on a message to one of their neighbors, it means that they wish for this information to spread. This agent, therefore, also wish for this information to spread to other neighbors.

and learned that the state is 0), they will have to fact-check their message. Furthermore, reliability is *public*: agents who receive messages from j will know that it has been fact-checked.

2.3 Summary of the game

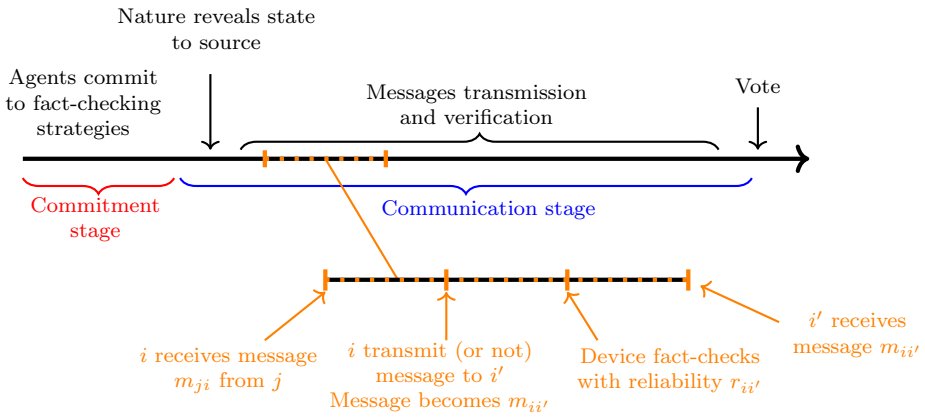

Figure 1: Timing of the game

I will refer to the game presented in this section as the **fact-checking game**. The fact-checking game comprises two stages. In the first stage, or **commitment** stage, agents commit to fact-checking. In the second stage or **communication** stage, agents transmit messages and vote. The subgame played in the communication stage essentially is the game presented in Bloch et al. (2018). The authors characterize several equilibria of this game in the corresponding paper, and I will rely on these results in my analysis. My contribution pertains to the new first stage, and as such, I will focus much of this paper analysis on this commitment stage.

The timing of the game is summarized on figure 2.3. Agents choose communication and fact-checking strategy in order to minimize cost of commitment in the first-stage and to maximize the probability that their preferred collective decision is implemented in the second stage. Writing u_{U_i} utility for unbiased agent i, and u_{B_j} for biased agent j, we can write utility functions as:

$$u_{U_i}(x,\theta) = -\lambda_U(x-\theta)^2 - c\sum_{i'\in N_i} r_{ii'}$$
$$u_{B_j}(x,\theta) = -\lambda_B(x-1)^2 - c\sum_{j'\in N_j} r_{jj'}.$$

I denote by $\boldsymbol{r} \in [0,1]^{|\mathcal{E}|}$ the (fact-checking) strategy profile of the first stage and with $\boldsymbol{\sigma}$ the strategy profile of the second stage[3]. The strategy profile of the whole game can therefore be written $(\boldsymbol{r}, \boldsymbol{\sigma})$.

[3]The second stage is a fully defined sequential game, so by strategy profile I mean a contingent plan of actions for each player.

2.4 Discussion

The fact-checking game relies on assumptions that may seem restrictive if interpreted literally. They are tools for tractability, and I argue that for a laxer interpretation of their meaning.

2.4.1 Message transmission

There are two potential roles for agents in the communication stage: source or messenger. The **source** is a unique agent appointed by Nature. This agent is informed about the true state, but their main power is not their knowledge. It rather is their capacity to create *any* message without constraints, *i.e.* to lie. **Messengers**, agents who are not the source, can only *transmit* messages they receive, without altering them. In section 5 some empirical situations where these assumptions might apply.

Source and messengers have different strategy spaces (the former can lie and not the later). This assumption is realistic only in specific contexts, such as the sharing of images or videos on social media. The development of new techniques to falsify images or videos such as *deepfake* makes such situations more likely to occur. Nevertheless, the scope of applications remains limited. I argue that it can reach more generality if interpreted as a *cognitive* statement on how agents update their beliefs. In this interpretation, agents act as if the "source" is the last person to have transformed the message - discarding all history before this last transformation. The critical feature is that upon receiving a message, one agent thinks that *someone* in the branch where the message came from is the source. The **random source** assumption ensures that agents do not automatically assume that the original source of a message is the person directly sending this message to them. The random source assumption states that before sending a signal, Nature chooses the source *uniformly at random*. It can be interpreted as agents having uniform priors about who is the source in the branch when they receive a message. In other words, agents think that it is equally likely that any agent in the branch where the message comes had the opportunity to transform it. In the presence of uncertainty over the source, agents may use this sort of rule of thumb. The specific form of *random source*, uniform distribution, is made for tractability: the results hold if priors over sources have another form.

2.4.2 Fact-checking

The most direct interpretation of fact-checking devices is an actual technology used by platform designer. Imagine an online platform where messages circulate on a network in the way described before. The *fact-checking device* is an option offered to the members of the platform when they form a new link. If they choose that option, any messages that the agent sends will be fact-checked by the platform. The reliability level an agent chose for this fact-checking is displayed publicly on his profile and is available for all to see (not only their neighbor).

One might wonder why agents would want the platform to police messages they send. For truth-seeker agents, this is obvious why: they use the fact-checking device to *fact-check*. That is, they buy a right to verify information. For biased agents, this is less clear. Choosing to be fact-checked more will increase their chances of being caught lying. On the other hand,

because this choice is *public*, it also improves the credibility of the messages they send. An agent who gets fact-checked with good reliability would be more likely to be believed, and can, therefore, exercise more influence. This trade-off between fewer opportunities to lie and more credibility is the core idea of Bayesian Persuasion. As Kamenica and Gentzkow [Kamenica & Gentzkow (2011)] shows, this trade-off is allowed by the commitment assumption. One exciting feature of the model is that - because information circulates on a network - the credibility of messages depends on the reliability of all agents that are upstream on a branch in the network. The reliability chosen by others will, therefore, determine the optimal reliability chosen by any given agent. As I will show in the paper, the tree nature of the network allows us to determine optimal reliabilities sequentially, starting from the end of the tree and going toward the center.

The idea of a fact-checking device implemented by a platform is convenient for understanding, but other interpretations might hold. Reliability might emerge from a reputation of integrity, for example. It might be that agents typically verify up to a certain degree r_{ij} information before transmitting it and that this "integrity" is publicly known (and it would be costly in the long term to deviate from it). The cost could be seen as the cost of information gathering or the education cost needed to assess the quality of information.

As a remark note that I do not allow for false positive in the definition of the fact-checking device. If the fact-checking device flags a message as false, it must be false. This is without loss of generality. Assume that false negative have a positive probability. In this case, agents who did not get information might fear that they did not receive a message because the fact-checking device gave a false negative. However, they also know that not receiving a message may happen if Nature did not send a message or someone blocked it. In this case, the posteriors, when not receiving a message, are surely less than the prior for $\theta = 1$: μ_0. This is also the case the no false-negative assumption[4]. Because $\mu_0 < \frac{1}{2}$, the presence of false negatives will not create new shifts in beliefs (unbiased agents not receiving messages will still vote for 0).

3 Full Communication compatibility

There exist a set of fact-checking strategies that leads agents to believe any messages they receive in the second stage. I call such set **Full Communication Compatible** strategies. This section shows the existence of this set and characterizes it.

3.1 Equilibrium Concept

In this paper, I ask if a designer can choose a cost for fact-checking such that communication and trust are entirely restored between agents in the second stage. **Trust** means that beliefs

[4]See footnote 15 in [Bloch et al. (2018)] for more details on updating when no message is received. The core idea is that because biased agents always block messages equal to 0, these messages are more likely to be blocked before reaching i than 1s. On the other hand, the absence of a signal from Nature or symmetric false-negative impacts the probability of 1 and 0, not reaching i in the same way. Hence, the absence of a message indicates that the state is more likely to be 0.

of any (unbiased) agent j upon receiving a message m_{ji} should shift in the direction of the message. Let $\mu_i(m_{ji})$ be the probability that $\theta = 1$ according to i's posterior beliefs. Trust implies that $\mu_i(m_{ji} = 1) \geq \frac{1}{2}$ and $\mu_i(m_{ji} = 0) \leq \frac{1}{2}$. I write $\boldsymbol{\mu}$ the vector of all agents' beliefs.

I use **Perfect Bayesian Equilibrium** (PBE) as an equilibrium concept. Borrowing terminology from Bloch et al. (2018), a situation where communication is ensured on all edges is referred to as a **Full Communication Equilibrium** (FCE).

Definition 1.A. *For a given (fixed) fact-checking profile $\boldsymbol{r}$, a set of strategies and beliefs $(\boldsymbol{\sigma}, \boldsymbol{\mu})$ is a **Full Communication Equilibrium of the second-stage** if $(\boldsymbol{\sigma}, \boldsymbol{\mu})$ forms a PBE in the subgame played in the second stage actions and:*

- *All unbiased sources create trustful messages (i.e. messages matching the signal send by Nature).*
- *All unbiased messengers transmit any message they received.*

I extend naturally this definition to characterize strategies of the entire game:

Definition 1.B. *A set of strategies and beliefs $\big((\boldsymbol{r}, \boldsymbol{\sigma}), \boldsymbol{\mu}\big)$ is a **Full Communication Equilibrium** if it is PBE of the fact-checking game and, given $\boldsymbol{r}$, the profile $(\boldsymbol{\sigma}, \boldsymbol{\mu})$ forms an full communication equilibrium of the second stage.*

Note that Full Communication Equilibria do not restrict "fact-checking" strategies $\boldsymbol{r}$. The question I want to ask is how $\boldsymbol{r}$ should be set in the first stage for the above strategies to hold as a Perfect Bayes Nash Equilibrium in the second stage. If such FCE exists it would mean that fact-checking commitment could potentially solve communication failures.

The key feature of Full Communication Equilibrium is that unbiased agents transmit any message they receive. Remember that from [Bloch et al. (2018)], unbiased agents only transmit messages if they believe it. For such a strategy to hold in a PBE, it must be that unbiased agents believe all messages they receive. In formal terms, it means that for any unbiased agent i, i's posterior upon receiving any message should shift in favor of that message[5]. More specifically, beliefs are consistent with Full Communication in the second stage if and only if $\mu_j(m_i = 1) \geq \frac{1}{2}$, $\mu_j(m_i = 0) \leq \frac{1}{2}$ and $\mu_j(m_i = \emptyset) \leq \frac{1}{2}$ for all $i \in \mathcal{N}_j$, for all $j \in \mathcal{U}$.

The condition $\mu_j(m_i = 0) \leq \frac{1}{2}$ arise naturally. Only unbiased sources have the incentive to send messages equal to 0; hence a message equal to 0 should always be true. Similarly, if j does not receive any message, it can be either because an agent blocked a 1 message, a fact-checking device caught a false 1 message, or because Nature did not send any signal in the first place. The first two events indicate that $\theta = 0$ is more likely, while the second does not inform on the state. Hence, surely $\mu_j(m_i = \emptyset) \leq \frac{1}{2}$. The condition $\mu_j(m_i = 1) \geq \frac{1}{2}$ is more complex as the credibility of $m_i = 1$ depends on i's fact-checking behavior as well as their place in the network. The remainder of this section characterizes conditions on fact-checking strategies to obtain $\mu_j(m_i = 1) \geq \frac{1}{2}$ when agents play Full Communication strategies in the second stage.

[5]One should not worry about biased agents "trust". Indeed because biased agents' payoff is independent of the state of the world, their strategies do not depend on their beliefs in equilibrium. Hence biased agents' beliefs are irrelevant for the analysis of communication.

3.2 Full Trust Equilibrium

Let $G_i(j)$ denote the subgraph upstream of an edge ji, *i.e.*, where a message going from j to i might have originated. Figure 2 illustrates this concept. $\mathcal{B}_i(j)$ is the set of biased agents in this subgraph, and $\mathcal{U}_i(j)$ the set of unbiased agents. I denote $S_i(j) = \mathcal{B}_i(j) \bigcup \mathcal{U}_i(j)$, the set of all agents (nodes) in the subgraph. Finally let $b_{S_i(j)} = \frac{|\mathcal{B}_i(j)|}{|S_i(j)|}$ be the proportion of biased agents in $G_i(j)$.

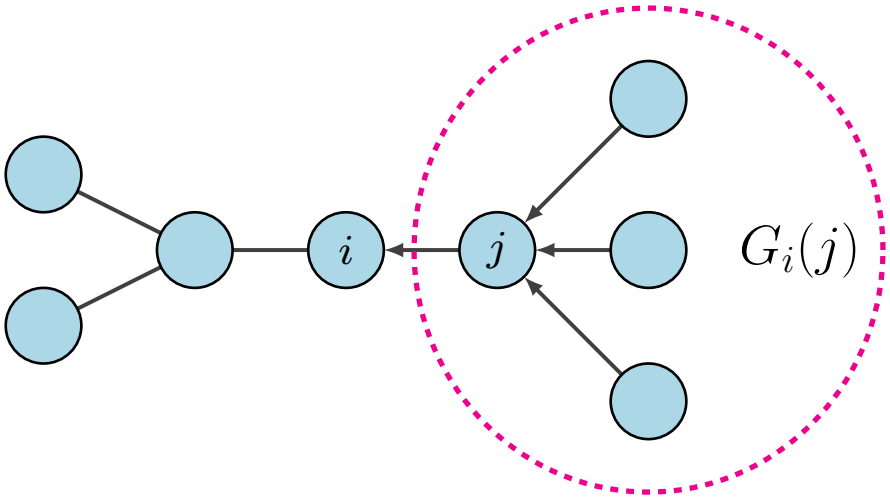

Figure 2: Definition of $G_i(j)$.

When $b_{S_i(j)}$ is very low, the probability that a biased agent is the source of a message flowing from j to i is very low. In some networks where biased agents are sparsely distributed, there exists Full Communication Equilibrium where no agents are fact-checking. Following Bloch et al. (2018), I call such an equilibrium a Full Trust Equilibrium.

Theorem 1. *(Adapted from Bloch et al. (2018)) A Full Communication Equilibrium where $r_{ij} = 0$ for all edges $ij \in \mathcal{E}$ (**Full Trust equilibrium**) exists if and only if for each unbiased agents i and each of its neighbors j:*

$$\frac{\mu_0}{1-\mu_0} \geq b_{S_i(j)}. \tag{1}$$

Proof. See Appendix in [Bloch et al. (2018)]. □

A Full Trust Equilibrium will occur if enough unbiased agents exist in all branches so that when a message circulates, the likelihood that it comes from an unbiased source is sufficiently high.

Example 1. *Consider the example in figure 3, with four agents: three unbiased and one biased. Suppose U_1 receives a message $m_{B_2} = 1$. U_1 posteriors will be computed with Bayes' rule:*

(a) Full Trust Equilibrium does hold

(b) Full Trust Equilibrium does not hold

Figure 3: Full Trust Equilibrium

$$\mu_{U_1}(m_{B_2}=1)=\frac{\mu_0\times P(m_{B_2}=1|\theta=1)}{\mu_0\times P(m_{B_2}=1|\theta=1)+(1-\mu_0)\times P(m_{B_2}=1|\theta=0)}.$$

To compute $P(m_{B_2}=1|\theta=1)$*, consider first what happens if* $\theta=1$*. Any agent designated as source have strong incentive to send a true message, including* B_2*. Furthermore, if* U_4 *or* U_3 *is the source and send a message* 1*,* B_2 *will have incentive to transmit this message to* U_1*. Hence,* $P(m_{B_2}=1|\theta=1)=1$*. Suppose now that* $\theta=0$*. If* U_4 *and* U_3 *are the source, they will send a message equal to* 0*, but then* B_2 *will block it. On the other hand, if* B_2 *is the source, they will send a message equal to* 1*. If* B_2 *doesn't fact-check, this message will always reach* U_1*. Hence when* $\theta=0$*,* U_1 *receives a message equal to* 1 *only if the source is biased:* $P(m_{B_2}=1|\theta=0)=b_{S_{U_1}(B_2)}=\frac{1}{3}$*. For* $\mu_0=0.3$*,* U_1 *forms the following posteriors:*

$$\mu_{U_1}(m_{B_2}=1)=\frac{\mu_0\times 1}{\mu_0\times 1+(1-\mu_0)\times\frac{1}{3}}\approx 0.53\geq\frac{1}{2}.$$

Hence, U_1 *will believe any message they receive, even without fact-checking. Because the situation is symmetric for other unbiased agents, we can conclude that a Full Communication Equilibrium exists, with no fact-checking.*

Consider now figure 3(b). U_1 *posterior can be computed using the same method. For* $\mu_0=0.3$*, we obtain:*

$$\mu_{U_1}(m_{B_2}=1)=\frac{\mu_0\times 1}{\mu_0\times 1+(1-\mu_0)\times\frac{1}{2}}$$
$$\Rightarrow\mu_{U_1}(m_{B_2}=1)\approx 0.46\leq\frac{1}{2}.$$

It means that, without fact-checking, U_1 *will not believe a message* m_{B_2} *received from* B_2*. It follows that a Full Trust Equilibrium with no fact-checking does not exist.*

In Figure 3(a), the low number of biased agents relative to unbiased agents plays in favor of trust. Indeed, messages are more likely to come from unbiased sources; hence they are trusted. Because U_1 is already trustful, there is no need for B_2 to use fact-checking to

persuade more, whatever the cost of fact-checking is. In Figure 3(b), on the other hand, the concentration of biased agents on the branch coming toward U_1 is too high. There are not enough unbiased agents on that branch to convince them that the message is trustworthy. The condition in Theorem 1 captures precisely this phenomenon.

In the next subsection, I show how B_2 can improve their credibility by fact-checking more. If B_2 fact-checks enough, U_1 will end up believing them, and trust will be restored.

3.3 Full-Communication Compatible Strategies

For which fact-checking strategies of B_2 does an FCE exist in the second stage? **Full Communication compatible** (FCC) strategies are such that, when played in the first stage, agents believe any message they receive in the second stage.

Definition 2. *A fact-checking strategy profile $\boldsymbol{r}$ is **Full-Communication compatible** if when $\boldsymbol{r}$ is played in the first stage, then a Full Communication Equilibrium of the second stage exists.*

Let's denote $\boldsymbol{r}_{\mathcal{U}} \in [0,1]^{|\mathcal{E}_{\mathcal{U}}|}$ the profile of unbiased agents fact-checking strategies. For each value of $\boldsymbol{r}_{\mathcal{U}}$, I construct a set of fact-checking strategies for biased agents ($\boldsymbol{r}_{\mathcal{B}} \in [0,1]^{|\mathcal{E}_{\mathcal{B}}|}$) that lead to a Full Communication Compatible profile.

3.3.1 No unbiased agents fact-check

To build intuition, I first characterize the set of Full Communication Compatible strategies in a simple case: when unbiased agents do not fact-check. This case illustrates the paper's core idea: fact-checking can be a persuasion device for biased agents. Consider the following example.

Example 2. *Consider a network of only two agents: U_1 (unbiased) and B_1 (biased). When B_1 is the source, this can be analyzed as a traditional Sender-Receiver problem.*

*Without fact-checking, **cheap-talk** equilibria will impede communication. U_1 knows that B_2 will send $m_{B_1} = 1$ for any signal they received from Nature. It follows that communication between the two will fail.*

With fact-checking, and more importantly commitment, B_1 is given the chance to persuade U_1. The posterior belief of U_1 after receiving $m_{B_2} = 1$ can be computed as:

$$\mu_{U_1}(m_{B_1} = 1) = \frac{P(m_{B_1} = 1|\theta = 1) \times \mu_0}{P(m_{B_1} = 1|\theta = 1) \times \mu_0 + P(m_{B_1} = 1|\theta = 0) \times (1 - \mu_0)}.$$

This problem is studied extensively by Kamenica and Genthworz in their seminal Bayesian persuasion paper [Kamenica & Gentzkow (2011)]. The paper shows that if B_1 can persuade that $P(m_{B_1} = 1|\theta = 0) \leq \frac{\mu_0}{1-\mu_0}$, then cheap talk can be avoided. Unfortunately, this cannot be achieved in traditional settings because, conditioning on receiving $\theta = 0$, $m_{B_1} = 1$ is always a dominant strategy. In [Kamenica & Gentzkow (2011)] a way around this issue is found by allowing agents to commit to certain messages when they receive certain signals from Nature.

The fact-checking game can be seen as a simple application of this persuasion setting. Let $\sigma_{B_1}(m_{B_1} = 1|\theta = 1)$ *be the sending strategy of* B_1 *and* $r_{B_1U_1}$ *be their fact-checking strategy.* U_1 *posteriors can be computed as:*

$$\mu_{U_1}(\theta = 1|m_{B_1} = 1) = \frac{\sigma_{B_1}(m_{B_1} = 1|\theta = 1) \times \mu_0}{\sigma_{B_1}(m_{B_1} = 1|\theta = 1) \times \mu_0 + (1 - r_{B_1U_1})\sigma_{B_1}(m_{B_1} = 1|\theta = 0) \times (1 - \mu_0)}.$$

In this simple example, the formula shows that a commitment over $r_{B_1U_1}$ *is functionnaly equivalent to a commitment over* $\sigma_{B_1}(m_{B_1} = 1|\theta = 0)$.

Example 2 shows that, in an elementary setting, *there is* a fact-checking strategy for a biased agent such that the receiver will believe the message. With my terminology, in the Sender-Receiver example, any fact-checking profile with $r_{B_1} \geq 1 - \frac{\mu_0}{1-\mu_0}$ is Full-Communication compatible. Can this be extended to more complex networks? To show it, I need to introduce some definitions.

Definition 3. *An edge ji is* ***problematic*** *if j is unbiased, i is biased and*

$$\frac{\mu_0}{1 - \mu_0} \leq b_{S_j(i)}.$$

An edge that is not problematic is an edge where trust is ensured even without fact-checking. One way to express Theorem 1 is to say that Full Trust Equilibrium holds if and only if there are no problematic edges in the network.

If an edge ij is *not* problematic, then any message going through that edge will be believed even with $r_{ij} = 0$. I show that if biased agents fact-check sufficiently, they can ensure trust on problematic edges. Example **??** illustrates how:

Example 3. *Consider the network depicted on figure* **??**. *The directed edges in red are problematic edges. Without any fact-checking, a message coming from* B_2 *will not be believed by* U_1 *nor by* U_5. *However, all messages are believed on any other edges on the network. We can use problematic edges to divide the network into three sub-networks (circled), each inside of which there is full trust. Consider the posteriors of* U_1 *after receiving* $m_{B_2} = 1$. *When receiving this message,* U_1 *knows that the source is someone in* $S_{U_1}(B_2)$ *(yellow and blue circle). More precisely, because of the (uniform) random source assumption,* U_1 *thinks that there is a probability* $b_{S_{U_1}(B_2)} = \frac{4}{7}$ *that the source is biased and a probability* $1 - b_{S_{U_1}(B_2)} = \frac{3}{7}$ *that the source is unbiased.* U_1 *also knows that biased sources sends* $m = 1$ *for any signal* θ, *and unbiased sources always sends truthful messages. Suppose all agents in* $S_{B_2}(U_1)$ *play Full communication equilibrium in the second stage. When a message originates in* $S_{B_2}(U_1)$, *the only problematic edge it goes through on its way to* U_1 *is* B_2U_1. *Hence, we get:*

$$P(m_{B_2U_1} = 1|\theta = 1) = \mu_0 \times [b_{S_{B_2}(U_1)}\sigma_B(1|\theta = 1) + (1 - b_{S_{B_2}(U_1)})\sigma_U(1|\theta = 1)],$$
$$P(m_{B_2U_1} = 1|\theta = 0) = (1 - \mu_0) \times [b_{S_{B_2}(U_1)}(1 - r_{B_2U_1})\sigma_B(1|\theta = 0) + (1 - b_{S_{B_2}(U_1)})\sigma_U(1|\theta = 0)].$$

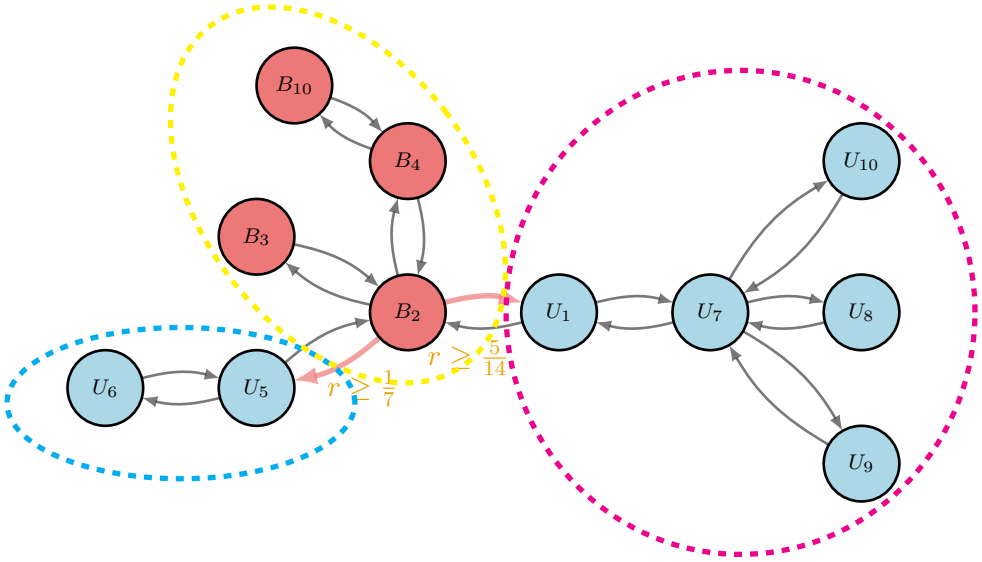

Figure 4: Full Communication Compatible strategies with problematic edges

The posterior of U_1 *can therefore be computed as*

$$\begin{aligned}\mu_{U_1}(\theta = 1|m_{B_2} = 1) &= \frac{1 \times \mu_0}{1 \times \mu_0 + [b_{S_{B_2}(U_1)} \times (1 - r_{B_2U_1}) \times 1 + \frac{2}{7} \times 0](1 - \mu_0)} \\ &= \frac{\mu_0}{\mu_0 + b_{S_{B_2}(U_1)}(1 - \mu_0)(1 - r_{B_2U_1})}.\end{aligned}$$

For any $r_{B_2U_1} \geq 1 - \frac{1}{b_{S_{B_2}(U_1)}} \frac{\mu_0}{(1-\mu_0)}$, *we will have* $\mu_{U_1}(\theta = 1|m = 1) \geq \frac{1}{2}$. *Similarly, it can be shown that any* $r_{B_2U_5} \geq 1 - \frac{1}{b_{S_{B_2}(U_5)}} \frac{\mu_0}{(1-\mu_0)}$ *is Full Communication Compatible.*

Therefore, any fact-checking profile where

$$\begin{aligned}r_{B_2U_5} &\geq 1 - \frac{1}{b_{S_{B_2}(U_1)}} \frac{\mu_0}{(1 - \mu_0)}, \\ r_{B_2U_5} &\geq 1 - \frac{1}{b_{S_{B_2}(U_5)}} \frac{\mu_0}{(1 - \mu_0)},\end{aligned}$$

and $r_{ij} = 0$ *on any other edge* ij *is Full Communication compatible.*

Biased agents involved in problematic edges can improve their credibility via *fact-checking* and thus ensure communication. The key difference with the simple sender-receiver case is that the amount of fact-checking necessary for them to be credible enough also depends on the proportion of biased agents on the branch they lie on.

In the network depicted in Figure 3, any message coming from any agents goes through only *one* problematic edge. In more general networks, there might be *chained* problematic edges. Consider the network depicted on Figure 5. If a message goes through a problematic edge (B_2, U_3), where it will be fact-checked, it is possible that it has already been through a problematic edge upward (B_4, U_1) where it was also fact-checked. Therefore, U_3 will base his beliefs updating on all the fact-checking that occurred upward of B_2.

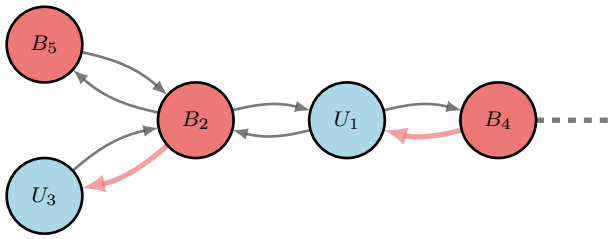

Figure 5: "Chained" problematic edges.

Example 4. *Consider the network depicted on figure 5. Suppose that on the right side of B_4 the network continues such that there is n agents in the network, hence $|S_{B_1}(B_4)| = n-4$ (that is the number of agents upstream of $B_4 - U_1$, B_4 included). Like in preceding example U_3 posteriors can be computed as:*

$$\mu_{U_3}(m_{B_2} = 1) = \frac{\mu_0}{\mu_0 + (1-\mu_0)P(m_{B_2} = 1|\theta = 0)}.$$

In this case however $P(m_{B_2} = 1|\theta = 0)$ not only depends on the fact-checking behavior of B_2 but also of B_4. Let S be an indice that denotes a set of agent where the source might be. $S = S_{U_1}(B_4)$ for example means that the source is someone in $S_{U_1}(B_4)$. With this indice, we can write:

$$P(m_{B_2} = 1|\theta = 0) = \sum_{s \in \{B_2, U_1, B_5, \} \cup S_{U_1}(B_4)} P(m_{B_2} = 1|\theta = 0 \cap s = S).$$

Assume that in the subnetwork $G_{U_3}(B_2)$ Full Communication compatible strategies are played (this will hold when we will be in equilibrium in the next section). All message will be transmitted, hence we can write:

$$P(m_{B_2} = 1|\theta = 0) = \frac{2}{n-1}(1 - r_{B_2 - U3}) + 0 + \frac{n-4}{n-1}(1 - r_{B_2 - U3})P(m_{B_2} = 1|\theta = 0).$$

Furthermore, if Full Communication compatible strategies are played, $\mu_{U_1}(m_{B_4} = 1) \geq \frac{1}{2}$. Using the formula for posteriors, this can be rewritten as: $P(m_{B_4} = 1|\theta = 0 \cap s = S_{U_1}(B_4)) \geq \frac{\mu_0}{1-\mu_0}$. It follows that

$$\mu_{U_3}(m_{B_2} = 1) \geq \frac{\mu_0 \times (n-1)}{(n-1)\mu_0 + (1 - r_{B_2 - U3})[2(1-\mu_0) + \mu_0(n-4)]}).$$

Therefore,

$$\mu_{U_3}(m_{B_2} = 1) \geq \frac{1}{2} \Longrightarrow r_{B_2 - U_3} \geq 1 - \frac{(n-1)\mu_0}{2(1-\mu_0) + \mu_0(n-4)}.$$

The key take-away from Example 4 is that, with chained problematic edges, we can use the fact that - in equilibrium - unbiased receiver posteriors should be greater than one half (otherwise they block) to compute the Full Communication compatible strategies down the branch. Therefore, Full Communication compatible on edge ij will depend on:

1. The number of agents upstream of other problematic edges in $S_i(j)$, let us denote this $\mathcal{N}_i^{chck}(j)$. In example 4 $\mathcal{N}_{B_2}^{chck}(U_3) = |S_{B_1}(B_4)| = n-4$.
2. The number of **biased** agents between ij and other problematic edges that are biased agents and might be the source of a message that goes through $B_2 - U_3$ such that this message is checked for the first time on that edge. Let us denote this $\mathcal{B}_i^{unchck}(j)$.

Generalizing example 4, we get the following proposition.

Proposition 1. *Suppose a unbiased fact-checking strategy profile where unbiased agents don't fact-check ($\boldsymbol{r}_{\mathcal{U}} = (0, \dots, 0)$). Suppose $\boldsymbol{r}_{\mathcal{B}}$ is biased strategy profile such that for all biased i and all $j \in \mathcal{N}_i$, we have for r_{ij}:*

$$r_{ij} = \begin{cases} \max(1 - |S_i(j)| \frac{\mu_0}{|\mathcal{B}_i^{unchk}(j)|(1-\mu_0) + |\mathcal{N}_i^{chk}(j)|\mu_0}, 0), & \text{if } j \text{ is unbiased,} \\ 0 & \text{if } j \text{ is biased.} \end{cases}$$

Then $(\boldsymbol{r}_{\mathcal{U}}, \boldsymbol{r}_{\mathcal{B}})$ is Full Communication Compatible.

Formal proof for Proposition 1 is available in Appendix A. This Proposition is silent on uniqueness. Consider the case in the network depicted in Figure 5 where B_5 perfectly fact-checks even if they are not in a problematic edge - it appears clearly that B_2 would require substantially less fact-checking to be convincing.

3.3.2 A general characterization of Full Communication compatibility

Full Communication compatible strategies follows through even if some unbiased agents fact-check. Consider the following example.

Example 5. *Consider again the network depicted in figure 5. Suppose now that U_1 perfectly fact-checks. Any false message coming from upstream of U_1 will be blocked. Hence, the only way to get a false message to reach B_2 is that the source is either B_2 or B_3. Hence, we can write:*

$$\mu_{U_3}(m_{B_2} = 1) = \frac{\mu_0}{\mu_0 + (1-\mu_0)(1 - r_{B_2 - U_1})\frac{2}{n-1}},$$

which gives the following Full Communication compatible strategies:

$$r_{B_2-U_1} \geq 1 - \frac{(n-1)\mu_0}{2(1-\mu_0)}.$$

Example 5 extends to other positive fact-checking (not necessarily perfect) from unbiased agents. In such cases, biased agents will continue to adjust their fact-checking behavior depending on the likeliness that the source is biased on the branch they lie on. Having

unbiased agents who fact-check (even imperfectly) will increase the likeliness that a source is biased (conditional on receiving a message) in such a branch because such unbiased agents will (imperfectly) block any false message coming from upstream. We, therefore, get the following Proposition:

Proposition 2. *For any unbiased fact-checking strategy profile $\boldsymbol{r}_{\mathcal{U}} = (r_{ij})_{i\in\mathcal{U}}$, there exists a **minimum trust-inducing** biased strategy profile (denoted $\boldsymbol{r}^t_{\mathcal{B}}$), such that for any $\boldsymbol{r}'_{\mathcal{B}}$ if $\boldsymbol{r}'_{\mathcal{B}} \geq \boldsymbol{r}^t_{\mathcal{B}}$, then $(\boldsymbol{r}'_{\mathcal{U}}, \boldsymbol{r}_{\mathcal{B}})$ is Full Communication compatible.*

The proof for Proposition 2 is available in the Appendix. The intuition is very similar to Proposition 1. I fix an unbiased profile $\boldsymbol{r}_{\mathcal{U}}$ and construct by iteration a Full Communication compatible biased profile $\boldsymbol{r}^t_{\mathcal{B}}$, starting from the ends of the tree and going toward the center. The iteration allows us to show that this profile is uniquely defined and only depends on $\boldsymbol{r}_{\mathcal{U}}$.

It will be convenient for future analysis to define minimum trust inducing strategies as a function of the underlying unbiased agents' strategies. Let

$$\boldsymbol{r}^t_{\mathcal{B}} : \begin{cases} [0,1]^{|\mathcal{E}_{\mathcal{U}}|} \to [0,1]^{|\mathcal{E}_{\mathcal{B}}|} \\ \boldsymbol{r}^t_{\mathcal{B}}(\boldsymbol{r}_{\mathcal{U}}) = \boldsymbol{r}_{\mathcal{B}}, \text{ with } \boldsymbol{r}_{\mathcal{B}} \text{ such that } (\boldsymbol{r}_{\mathcal{U}}, \boldsymbol{r'}_{\mathcal{B}}) \text{ is FCC for any } \boldsymbol{r'}_{\mathcal{B}} \geq \boldsymbol{r}_{\mathcal{B}}, \end{cases}$$

be a function that associates each $\boldsymbol{r}_{\mathcal{U}}$ to the corresponding biased minimum trust inducing strategy. This function will prove very useful to determine the best-responses of biased agents.

4 Implementing Full Communication

I identify a range of cost for fact-checking for which Full Communication equilibrium can be implemented.

4.1 Free fact-checking

Consider first the case $c = 0$. Theorem 2 states that free-fact checking leads to perfect fact-checking from unbiased agents.

Theorem 2. *If $c = 0$, then for all positive λ_U, λ_B, there exist a Full Communication equilibrium where unbiased agents perfectly fact-check all there edges ($r_{ij} = 1$ for all $i \in \mathcal{U}$ and $j \in N_i$).*

A formal proof of Theorem 2 is available in the Appendix. With free fact-checking, unbiased agents perfectly fact-check to learn the state. Biased agents, on the other hand, fact-check their minimum trust inducing strategy, which is very low when unbiased agents perfectly fact-check because there already is a large amount of trust in the network. Biased agents minimize their fact-checking behavior because they face an indirect cost for fact-checking: more fact-checking means that they are more likely to be caught lying.

4.2 Best-responses

The existence of a Full Communication Equilibrium will depend on the willingness of biased agents to fact-check sufficiently to keep their credibility or not. If unbiased agents do not fact-check a lot, biased agents will need to fact-check more to ensure credibility and hence be more sensitive to the cost.

Agents' benefits from fact-checking depend on two conditions: their motivation λ and their position in the network. The more motivated agents are, the more they are willing to pay for fact-checking if it helps them fulfill their goal. Agents' position in the network, on the other hand, plays an important role because it determines how many other agents will be affected by their fact-checking behavior. Let's denote $g_{ji}(G, \boldsymbol{r}_{\mathcal{U}})$ the expected proportion of agents a biased agent j will convince if they obtain from their neighbor i to transmit their message. $g_{ji}(G, \boldsymbol{r}_{\mathcal{U}})$ corresponds to the benefit j get from fact-checking at least their minimum trust-inducing strategy on ji (if they do so i will believe, hence transmit, their message).

Proposition 3. *For any (λ_B, λ_U) and a fixed unbiased fact-checking profile $\boldsymbol{r}_{\mathcal{U}}$, there exist a $\bar{c}_{\boldsymbol{r}_{\mathcal{U}}} = \lambda_B \max_{j\in\mathcal{B}}(\max_{i\in\mathcal{N}_j}(g_{ji}(G, \boldsymbol{r}_{\mathcal{U}}))$ that depends on λ_B and $\boldsymbol{r}_U$ such that for any $c \leq \bar{c}_{\boldsymbol{r}_{\mathcal{U}}}$, biased agents best response to $\boldsymbol{r}_{\mathcal{U}}$ with their minimum trust inducing strategy.*

Proposition 3 states that, if the cost of fact-checking is under a specific cost $\bar{c}$, biased agents best-response to any unbiased fact-checking profile $\boldsymbol{r}_{\mathcal{U}}$ is their corresponding minimum trust-inducing strategies (defined in Proposition 2).

For unbiased agents, $\frac{|U_i(j)|}{N}$ captures the maximum proportion of agents they can affect by fact-checking ij. Because unbiased agents play in dominant strategies, this is sufficient to capture how network position affect their incentives.

Proposition 4. *For any (λ_u, λ_B), there exist a cost $\underline{c} = \lambda_U \max_{i\in\mathcal{U}}(\max_{j\in N_i}(\frac{|U_i(j)|-1}{N}))$, such that if $c \geq \underline{c}$, no unbiased agents fact-check (in dominant strategy).*

Proposition 4 implies that there exists a maximal cost, which is a function of λ_B and G, for which unbiased agents are ready to fact-check in dominant strategies. Whatever the behavior of biased agents is, above $\underline{c}$, unbiased agents will not fact-check.

4.3 Full Communication Equilibrium when unbiased agent don't fact-checking

A direct application of best-responses mechanics can gives us a simple range of cost where full communication can be implemented. Let us denote

$$g = \frac{\max_{i\in\mathcal{U}}(\max_{j\in N_i}(\frac{|U_ij|-1}{N})}{\max_{j\in\mathcal{B}}(\max_{i\in\mathcal{N}_j}(g_{ji}(G, \boldsymbol{r}_{\mathcal{U}} = \mathbf{0}))}.$$

When the network is fixed, g is a constant. It captures the ratio between unbiased "risk" of not fact-checking (maximum number of agents misled) and biased opportunity of fact-checking. Observe that

$$\frac{\lambda_B}{\lambda_U} \geq g \iff \underline{c} \leq c_{\mathbf{0}}.$$

The following theorem then follows from a direct application of Proposition 4 and 3.

Theorem 3. *For all λ_U, λ_B, such that $\frac{\lambda_U}{\lambda_B} \geq g$, there exists $\bar{c}$, $\underline{c}$, such that for all $c \in [\underline{c}, \bar{c}]$, there exists a Full Communication equilibrium in the fact-checking game where:*

- *Unbiased agents don't fact-check.*
- *Biased agents involved on problematic edges fact check with reliability:*

$$r_{ij} = \max(1 - |S_i(j)| \frac{\mu_0}{|\mathcal{B}_i^{unchk}(j)|(1-\mu_0) + |\mathcal{N}_i^{chk}(j)|\mu_0}, 0).$$

A proof for 4 is available in the Appendix. The formula for biased fact-checking is derived from Proposition 2. The theorem shows that when the cost of fact-checking is very high, and biased agents are more motivated than unbiased agents, the *burden of proof* - the amount of fact-checking needed to ensure Full Communication - is carried by *biased* agents. One could think that fact-checking is to be used by truth-seeker agents, especially if the cost is high. Here is a situation where fact-checking is used entirely as a device of *persuasion* and not information gathering.

4.4 General conditions for the existence of Full Communication equilibrium

With a slightly more restrictive condition on preferences, we can get a larger set of cost to implement FCE. Denote

$$\tilde{g} = \frac{\max_{i \in \mathcal{U}}(\max_{j \in N_i}(\frac{|U_i(j)|-1}{N})}{\min_{\boldsymbol{r}_{\mathcal{U}}}(\max_{j \in \mathcal{B}}(\max_{i \in \mathcal{N}_j}(g_{ji}(G, \boldsymbol{r}_{\mathcal{U}}))))}.$$

Theorem 4. *For all λ_U, λ_B such that $\frac{\lambda_U}{\lambda_B} \geq \tilde{g}$, there is a $\bar{c}$, such that for all $c < \bar{c}$ there exists a Full Communication equilibrium.*

We know that if $c < \bar{c}$, biased agents will respond with full communication compatible strategies to any strategy played by unbiased agents. It is not clear however what the best response from unbiased agents to these strategies is. The proof of Theorem 4 shows that, as long as c is under the threshold, the best-response correspondence has a fixed point.

5 Discussion

In this subsection, I discuss some implications and limitations of the previous results.

5.1 Relaxing preferences restriction

The results above relied on assumptions on the motivation ratio $\frac{\lambda_B}{\lambda_U}$, which ensured an appropriate ordering of cost cutoffs for unbiased and biased agents. What happens to equilibrium behavior if we relax these assumptions? The existence of Full Communication equilibrium will depend on how cost c compares to the different cutoffs determined in Proposition 3 and 4. More precisely, the range of c, where Full Communication exists, will depend on the ordering of these cutoffs. This ordering itself will depend on agents' motivation and the network structure. Consider a situation where $\tilde{g} > \frac{\lambda_B}{\lambda_U}$. In that case, $\underline{c} > c_{\boldsymbol{r}_\mathcal{U}}$ for any $\boldsymbol{r}_\mathcal{U}$. When biased agents have little motivation compared to unbiased agents, they will stop fact-checking at a lower cost than unbiased agents. In this case, Full Communication is guaranteed only below $\tilde{c} = \min_{\boldsymbol{r}_\mathcal{U}} c_{\boldsymbol{r}_\mathcal{U}}$; above that threshold, biased agents are not ready to play the Full Communication compatible strategies.

5.2 Who carries the burden of proof?

Theorem 3 attests that a platform designer can choose between a range of costs to reach full communication at equilibrium. Does it mean that any implementation is equivalent?

In most networks, a certain amount of fact-checking is necessary to ensure communication. Not all agents typically fact-check however, the fact-checking of *some* agents might be sufficient to restore trust. I say that these agents carry the **burden of proof**. According to Theorem 2, if the cost is 0, unbiased agents fact-check perfectly, and biased agents fact-check very little: unbiased agents carry the burden of proof. On the other hand, according to Theorem 3, if the cost is sufficiently high *only* biased agents fact-check. The burden of proof relies completely on biased agents.

Such results imply that platform designer have some flexibility in how to implement full communication. Indeed, the cost of fact-checking will impact who carries the burden of proof but also how often messages are fact-checked - impacting the overall intensity of communication. An interesting direction for future research would be to carry comparative statics exercises to identify how a designer can affect optimize other objectives in choosing the cost of fact-checking.

5.3 Assumptions and interpretation

My paper builds on a voluntarily simplistic model and relies on restrictive assumptions that I would like to discuss here.

First, my model makes the critical assumptions that the primary source of information is unknown and that agents transmit it without modification. While some examples of social media communication do not match these assumptions (an article shared on Facebook as a source attached, for example), the empirical literature shows that many Internet communications display such features. Cagé, Hervé & Viaud (2020), for example, showed that about 50% of information on French media websites were simple copy-paste of an original piece, with the source mentioned in only 5% of cases. This number is probably much higher for social media, where information is informal. This contrasts heavily with traditional media, whose business model heavily relied on reputation mechanisms that somehow safeguard in-

formation quality [Gentzkow & Shapiro (2006)]. Several studies show how the lack of such safeguards, notably the mention of an information source, has led to an increase in the spread of false information [Allcott & Gentzkow (2017)].

Another crucial assumption that I make is that there are biased agents and that their type is common knowledge. I will argue that the critical feature of types is not that an agent is biased, but that others think this agent is biased. Suppose that Bill Gates genuinely wants to share useful health information on social platforms. Several fake news narratives claimed that Bill Gates was involved in the covid-19 crisis [Roose (2020)]. He will likely be considered as biased by some users and not be believed. In such a situation, it is realistic to think that Bill Gates might want to fact-check publicly the messages he sends to entice others to trust him.

Such interpretation may not seem applicable to many examples if one has in mind social media heavily studied by economists like Twitter. Indeed, on Twitter, users are exposed to many tweets from other users they do not know. In such cases, when types are not public, agents might signal their types when fact-checking, which introduces a new layer of complexity to the model. My interest is in different cases where the agents have a long-standing opportunity to observe each other and know each other's biases, whereas the information does not have a specific identifiable source. The instant messaging application *WhatsApp* is a good example of such a case. *WhatsApp* role in disseminating misinformation is more and more recognized [Manjoo (2018)]. Examples include its role in spreading conspiracy theories in the 2018 Brazilian election [Isaac & Roose (2018)] or how it fueled deadly mob violence in India [Goel, Raj & Ravichandran (2018)]. The private (and encrypted) nature of communication on WhatApps render the identification of primary sources of misinformation very difficult - even for the platform itself [Tyagi, Miers & Ristenpart (2019)]. Researchers studying the role of WhatApps during the Brazilian election showed that most of the misinformation was spread through images and videos that users rarely altered [Resende, Melo, Sousa, Messias, Vasconcelos, Almeida & Benevenuto (2019)]. Finally, *WhatApp* is characterized by a network of strongly personalized relationships marked by repeated communication [O'Hara, Massimi, Harper, Rubens & Morris (2014)]. In such relationships, individual biases are likely to be known.

6 Conclusion

The spread of fake news throughout the coronavirus crisis has revealed that misinformation does not only harm by disseminating false beliefs but also by barring useful information to reach the population. Medical institutions and professionals had trouble broadcasting their message and recommendations due to a widespread lack of trust on social media. In this paper, I make a case for **sender-driven fact-checking** as a tool to restore trust and solve communication failures in social networks. I use a simple tractable model developed in [Bloch et al. (2018)] where agents strategically transmit a piece of information in a network. In this setting, lack of trust prevents the diffusion of useful information in some parts of the graph. I extend this setting by allowing agents to commit to fact-check any messages they send up to the reliability of their choice. I show that there always exists a set of trust-inducing fact-checking behavior that solves communication failures. Then, I establish

that for an appropriate cost of fact-checking agents will choose trust-inducing fact-checking in equilibrium. Therefore, platform designers can use such sender-driven fact-checking to restore trust and improve communication on social media. My paper also show that when the cost of fact-checking is low, unbiased agents carry the "burden of proof" that restores trust, i.e., fact-checks. On the other hand, when the cost is high, biased agents carry this burden.

An important question that I was not able to tackle in this master thesis pertains to the welfare implications of fact-checking. Are agents better off with fact-checking than when communication failures remain? If they are, which types of agents benefit from it? Is one type better off than the other? The response to this question is not straightforward. The situation with and without fact-checking both faces trade-off. Without fact-checking, any suspect information is filtered out, at the risk of losing some useful information. With fact-checking, any right information will reach be transmitted; there is a risk, however, that some wrong information flows too.

References

Allcott, H. & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Ambrus, A., Azevedo, E. M., & Kamada, Y. (2013). Hierarchical cheap talk. *Theoretical Economics*, *8*(1), 233–261.

Anderlini, L., Gerardi, D., & Lagunoff, R. (2012). Communication and Learning. *Review of Economic Studies*, *79*(2), 419–450.

Bloch, F., Demange, G., & Kranton, R. (2018). Rumors and Social Networks. *International Economic Review*, *59*(2), 421–448.

Cagé, J., Hervé, N., & Viaud, M.-L. (2020). The Production of Information in an Online World. *The Review of Economic Studies*, *87*(5), 2126–2164.

Calvó-Armengol, A., Martí, J. d., & Prat, A. (2015). Communication and influence. *Theoretical Economics*, *10*(2), 649–690.

Candogan, O. (2019). Persuasion in Networks: Public Signals and k-Cores. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, (pp. 133–134)., Phoenix, AZ, USA. Association for Computing Machinery.

Crawford, V. P. & Sobel, J. (1982). Strategic Information Transmission. *Econometrica*, *50*(6), 1431–1451.

Egorov, G. & Sonin, K. (2019). Persuasion on Networks. SSRN Scholarly Paper ID 3375521, Social Science Research Network, Rochester, NY.

Fudenberg, D. & Tirole, J. (1991). *Game Theory* (1 edition ed.). Cambridge, Mass: The MIT Press.

Galeotti, A., Ghiglino, C., & Squintani, F. (2013). Strategic information transmission networks. *Journal of Economic Theory*, *148*(5), 1751–1769.

Gentzkow, M. & Shapiro, J. (2006). Media Bias and Reputation. *Journal of Political Economy*, *114*(2), 280–316.

Gieczewski, G. (2020). Verifiable Communication on Networks. Working Paper.

Goel, V., Raj, S., & Ravichandran, P. (2018). How WhatsApp Leads Mobs to Murder in India (Published 2018). *The New York Times*.

Green, J. R. & Stokey, N. L. (2007). A two-person game of information transmission. *Journal of Economic Theory*, *135*(1), 90–104.

Hagenbach, J. & Koessler, F. (2010). Strategic Communication Networks. *The Review of Economic Studies*, *77*(3), 1072–1099.

Isaac, M. & Roose, K. (2018). Disinformation Spreads on WhatsApp Ahead of Brazilian Election. *The New York Times*.

Kamenica, E. & Gentzkow, M. (2011). Bayesian Persuasion. *American Economic Review*, (6), 2590–2615.

Katz, E. & Lazarsfeld, P. F. (1966). *Personal Influence, the Part Played by People in the Flow of Mass Communications*. Transaction Publishers.

Kreps, S. E. & Kriner, D. (2020). Medical Misinformation in the COVID-19 Pandemic. SSRN Scholarly Paper ID 3624510, Social Science Research Network, Rochester, NY.

Manjoo, F. (2018). The Problem With Fixing WhatsApp? Human Nature Might Get in the Way (Published 2018). *The New York Times*.

O'Hara, K. P., Massimi, M., Harper, R., Rubens, S., & Morris, J. (2014). Everyday dwelling with WhatsApp. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, CSCW '14, (pp. 1131–1143)., New York, NY, USA. Association for Computing Machinery.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., & Benevenuto, F. (2019). (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *The World Wide Web Conference*, WWW '19, (pp. 818–828)., New York, NY, USA. Association for Computing Machinery.

Roose, K. (2020). Get Ready for a Vaccine Information War. *The New York Times*.

Tyagi, N., Miers, I., & Ristenpart, T. (2019). Traceback for End-to-End Encrypted Messaging. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, (pp. 413–430)., New York, NY, USA. Association for Computing Machinery.

Appendix

A.1. Full Communication compatibility

Proof of Proposition 1

Proof. This proof works by construction. For each edge ji, where j is biased, I define a minimum trust inducing fact-checking value $\boldsymbol{r}^t_{ji}(0) = \max(1 - |S_i(j)|\frac{\mu_0}{|\mathcal{B}_i^{unchk}(j)|(1-\mu_0)+|\mathcal{N}_i^{chk}(j)|\mu_0}, 0)$. I show that, when unbiased agents don't fact-check, this value is such that i's posterior (when receiving a message equal to 1) are equal to $\frac{1}{2}$ when computed with Bayes rule.

Fix a profile of unbiased fact-checking strategies $\boldsymbol{r_{\mathcal{U}}}$. Let's express $\mu_i(m_j = 1)$ as a function of fact-checking strategies. Applying Bayes rule we have:

$$\mu_i(m_j = 1) \geq \frac{1}{2} \quad \iff \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-r_{ij})P(m_j^{out} = 1|\theta = 0)} \geq \frac{1}{2},$$

where $P^{out}(m_j = 1|\theta = 0)$ is the probability that j sends a message equal to 1 to i when the state is 0. The condition $S_i(j)$ is here to remind that in such a case, i knows that the source is someone in $S_i(j)$ (it may be j themeself or someone else).

Observe that we have:

$$\mu_i(m_j = 1) \geq \frac{1}{2} \iff r_{ji} \geq 1 - \frac{\mu_0}{(1-\mu_0)P(m_j = 1|\theta = 0)}.$$

This proves that for all biased agents j and given others' strategies $\boldsymbol{r}_{-j}$, there exist a minimum trust inducing value $\boldsymbol{r}^t_{ji} = 1 - \frac{\mu_0}{(1-\mu_0)P(m_j=1|\theta=0)}$ such that i "believes" j.

Assume that unbiased agents don't fact-check ($\boldsymbol{r}_{\mathcal{U}} = (0, ...0)$). For any edge ji denote k the number of biased agents in $S_i(j)$. I will use an iteration on k to show that,

$$\boldsymbol{r}^t_{ji}(0) = 1 - \frac{\mu_0}{(1-\mu_0)P(m_j = 1|\theta = 0)} = \max(1 - |S_i(j)|\frac{\mu_0}{|\mathcal{B}^{unchk}_i(j)|(1-\mu_0) + |\mathcal{N}^{chk}_i(j)|\mu_0}, 0),$$

Initialization.

Take an edge ji such that there is 1 agents in $S_i(j)$. Then by definition j is the only biased agents in $S_i(j)$. Remember that $P(m^{out}_j = 1|\theta = 0)$ denotes the probability that j emits a message equal to 1 *before* fact-checking. Furthermore observe that agents different from j in $S_i(j)$ are all unbiased, therefore they don't fact-check. We can therefore write,

$$P(m^{out}_j = 1|\theta = 0) = \frac{1}{|S_i(j)|}.$$

Following definitions: $|\mathcal{B}^{unchk}_i(j)| = 1$ and $|\mathcal{N}^{chk}_i(j)| = 0$. Hence, we directly have:

$$r^t_{ij}(0) = \max(1 - |S_i(j)|\frac{\mu_0}{|\mathcal{B}^{unchk}_i(j)|(1-\mu_0) + |\mathcal{N}^{chk}_i(j)|\mu_0}, 0).$$

Iteration.

Suppose a fixed k. On all edges $j'i'$ that have k or less biased agents in $S_{i'}(j')$, suppose $\boldsymbol{r}^t_{j'i'}(0) = \max(1 - |S_i(j)|\frac{\mu_0}{|\mathcal{B}^{unchk}_i(j)|(1-\mu_0)+|\mathcal{N}^{chk}_i(j)|\mu_0}, 0)$. Suppose furthermore that such edges play there minimum trust inducing strategy $\boldsymbol{r}^t_{j'i'}$.

Consider on edge ji such that there is $k+1$ biased agents in $S_i(j)$. $P(m_j = 1|\theta = 0)$ depends on the strategies of biased and unbiased agents in $S_i(j)$. We can write:

$$P_{\boldsymbol{r}_{\mathcal{U}}}(m^{out}_j = 1|\theta = 0) = \sum_{i' \in \mathcal{B}_i(j)} \frac{1}{|S_i(j)|} \prod_{mn \in i' \to j} (1 - r_{mn}),$$

where $i' \to j$ is the set of all the edges (denoted by the indices mn) in the path from i' to j. If the state is 0 (as we assume it is), j can transmit a message equal to 1 only if she gets

one. This occurs if only if a biased agent is the *source*. For j to receive a message emitted by such a source, it must be that all fact-checking devices between i' and j fails, which is given by the probability $\prod_{mn \in i' \to j}(1 - r_{mn})$ ($\prod$ is the product operator).

Assuming that unbiased agents don't fact-check we know that all $(1 - r_{mn})$ where m is unbiased will be equal to one. Furthermore, assuming that biased agents in $S_i(j)$ play their minimum fact-checking strategy, we know that if r is biased, then $1 - r_{mn} = 1 - r^t_{mn}(0)$. Following definition, observe that if $m \in \mathcal{B}^{unchk}_i(j)$, then $r^t_{mn}(0) = 0$. Hence, we can make this first simplification:

$$P(m^{out}_j = 1|\theta = 0) = \frac{\mathcal{B}^{unchck}}{|S_i(j)|} + \sum_{i' \in \mathcal{B}_i(j) \setminus \mathcal{B}^{unchck}} \frac{1}{|S_i(j)|} \prod_{mn \in i' \to j} (1 - r_{mn}).$$

Consider now the second term of this expression. Observe that each edge $j'i' \in \mathcal{N}^{chck}$, we have $P(m^{out}(j') = 1|\theta = 0 \cap \mathbb{S}'_i(j')) = \sum_{i' \in \mathcal{B}'_i(j')} \frac{1}{|S'_i(j')|} \prod_{mn \in i' \to j}(1 - r_{mn})$.

We can therefore write:

$$P(m^{out}_j = 1|\theta = 0) = \frac{\mathcal{B}^{unchck}}{|S_i(j)|} + \sum_{i' \in \mathcal{N}^{chck}} \frac{1}{|S_i(j)|} P(m^{out}(j') = 1|\theta = 0).$$

Observe furthermore that from the iteration hypothesis, for all biased j' in $\mathcal{N}^{chck}$, we have:

$$P(m^{out}(j') = 1|\theta = 0) = \frac{\mu_0}{1 - \mu_0}.$$

Plugging this into the previous equation, we obtain:

$$P(m^{out}_j = 1|\theta = 0) = \frac{\mathcal{B}^{unchck}}{|S_i(j)|} + \sum_{i' \in \mathcal{N}^{chck}} \frac{1}{|S_i(j)|} \frac{\mu_0}{(1 - \mu_0},$$

or,

$$r^t_{ij}(0) = 1 - |S_i(j)| \frac{\mu_0}{|\mathcal{B}^{unchk}_i(j)|(1 - \mu_0) + |\mathcal{N}^{chk}_i(j)|\mu_0}$$

Because $|S_i(j)| \frac{\mu_0}{\mathcal{B}^{unchk}(1-\mu_0)+\mathcal{N}^{chk}\mu_0}$ is not necessarily non negative, we can write $r^t_{ij}(0) = \max(1 - |S_i(j)| \frac{\mu_0}{\mathcal{B}^{unchk}(1-\mu_0)+\mathcal{N}^{chk}\mu_0}, 0)$.

This iteration stops when $r^t_{ji}(0)$ has been defined for all biased agents' edges.

□

Proof of Proposition 2

Proof. This proof works by construction. For each edge ji, where j is biased, I construct a function $\boldsymbol{r}^t_{ji}(.)$, that only depends on $\boldsymbol{r}_{\mathcal{U}}$. This function is such that i's posterior (when receiving a message equal to 1) are equal to $\frac{1}{2}$ when computed with Bayes rule. *A priori*, i's posterior depends on the whole strategy profile of agents upstream of ji (*i.e.* in the graph $G_i(j)$). However, if we assume that minimum trust inducing strategies are played in this subgraph, these posteriors only depends on r_{ji} and $\boldsymbol{r}_{\mathcal{U}}$. From this observation, I use an

iteration that starts at the end of the graph and progressively goes upstream to show that $\boldsymbol{r}^t_{ji}(\boldsymbol{r}_{\mathcal{U}})$ is well defined for all j.

Fix a profile of unbiased fact-checking strategies $\boldsymbol{r}_{\boldsymbol{\mathcal{U}}}$. Let's express $\mu_i(m_j = 1)$ as a function of fact-checking strategies. Applying Bayes rule we have:

$$\mu_i(m_j = 1) \geq \frac{1}{2} \quad \iff \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-r_{ij})P(m_j^{out} = 1|\theta = 0)} \geq \frac{1}{2},$$

where $P^{out}(m_j = 1|\theta = 0)$ is the probability that j sends a message equal to 1 to i when the state is 0. In such a case, i knows that the source is someone in $S_i(j)$ (it may be j them-self or someone else).

Observe that we have:

$$\mu_i(m_j = 1) \geq \frac{1}{2} \iff r_{ji} \geq 1 - \frac{\mu_0}{(1-\mu_0)P(m_j = 1|\theta = 0)}.$$

This proves that for all biased agents j and given others' strategies $\boldsymbol{r}_{-j}$, there exist a minimum trust inducing value $\boldsymbol{r}^t_{ji} = 1 - \frac{\mu_0}{(1-\mu_0)P(m_j=1|\theta=0)}$ such that i "believes" j.

To complete the proof, we need to show one additional feature. In principle one biased agent's minimum trust inducing strategy depends on others biased agents strategy (via $P(m_j = 1|\theta = 0)$). Consider two biased agents j and j'. It is not guaranteed that r^t_{ji} and $r^t_{j'i'}$ settles on a fixed point. Because of the tree nature of the network, such loops will not exist, however. We can define the profile $\boldsymbol{r}^t_{\mathcal{B}}$ iteratively starting at the end of the tree and going toward the center. We will end up with a profile $\boldsymbol{r}^t_{\mathcal{B}}$ that only depends on unbiased agents fact-checking $\boldsymbol{r}^t_{\mathcal{U}}$. Let's prove it formally with an iteration on k, the number of biased agents upstream of a given edge ji (*i.e.* in $S_i(j)$).

Initialization.

Take an edge ji such that there is 1 agents in $S_i(j)$. Then by definition there j is the only biased agents in $S_i(j)$. Remember that $P(m_j^{out} = 1|\theta = 0)$ denotes the probability that j emits a message equal to 1 *before* herself has fact-checked. Hence, $P(m_j^{out} = 1|\theta = 0)$ only depends on the fact-checking behavior of unbiased agents $\boldsymbol{r}_{\mathcal{U}}$, which implies that $\boldsymbol{r}^t_{ji}$ only depends on $\boldsymbol{r}_{\mathcal{U}}$.

Iteration.

Suppose for a fixed k that for all edges $j'i'$ that have k or less biased agents in $S_{i'}(j')$, $\boldsymbol{r}^t_{j'i'}$ only depends on $\boldsymbol{r}_{\mathcal{U}}$. Suppose furthermore that such edges play there minimum trust inducing strategy $\boldsymbol{r}_{j'i'}$.

Consider on edge ji such that there is $k+1$ biased agents in $S_i(j)$. $P(m_j = 1|\theta = 0)$ depends on the strategies of biased and unbiased agents in $S_i(j)$. Consider any edge $j'i'$

in $G_i(j)$ such that j' is biased. By the hypothesis, there are at most k biased agents in $S_{i'}(j')$. Hence, $P(m(j') = 1|\theta = 0)$ only depends on unbiased strategies. Furthermore, j' minimum trust inducing strategy $\boldsymbol{r}^t_{j'i'}$ also only depends on $\boldsymbol{r}_{\mathcal{U}}$ (by hypothesis). Directly $P(m_j = 1|\theta = 0)$ (hence $\boldsymbol{r}^t_{ji}$) only depends on $\boldsymbol{r}_{\mathcal{U}}$.

The iteration stops when r^t_{ji} has been defined for all biased agents' edges.

Final conclusion

In conclusion, we can define a unique profile $\boldsymbol{r}^t_{\mathcal{B}}(\boldsymbol{r}_{\mathcal{U}}$ that depends only on unbiased agents strategy such that for all edge ji where j is biased $\mu_i(m_j = 1) \geq \frac{1}{2}$. By definition of Full Communication, this implies that the profile $(\boldsymbol{r}^t_{\mathcal{B}}(\boldsymbol{r}_{\mathcal{U}}), \boldsymbol{r}_{\mathcal{U}})$ is Full Communication compatible. □

A.2. Best responses

Proof of Lemma 4 To prove lemma 4, I first need to prove lemma 1 that identifies dominant strategies for unbiased agents.

Lemma 1. *Assuming all agents play Full Communication strategies in the second stage, if i is an unbiased agent then either $r_{ij} = 0$ or $r_{ij} = 1$ dominates all other strategies.*

Proof. In this proof, I first express the expected utility of any unbiased agent i as a function of their fact-checking behavior. This function is linear in i's strategies, implying that the maximum excepted utility is obtained either playing 0 or 1.

Step 1: Express unbiased agents payoff as a function of their fact-checking behavior.

Take i an unbiased agent and fix the profile of other players fact-checking strategies, that is $\boldsymbol{r}_{-i}$. Suppose that all agents (including i) play Full Communication strategies in the second stage.

I want to express the ex-ante utility of i as a function of their fact-checking strategies, that is $\boldsymbol{r}_i$, the vector of fact-checking from i on all edges they are connected to: $\boldsymbol{r}_i = (r_{ij})_{j \in N_i}$. I write this function as $U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i})$, where $\boldsymbol{r}_{-i}$ is fixed:

$$U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i}) = E(u_i(x, \theta)|(\boldsymbol{r}_i, \boldsymbol{r}_{-i})) = -\lambda_U E((x - \theta)^2|(\boldsymbol{r}_i, \boldsymbol{r}_{-i})) - c \sum_{j \in N_i} r_{ij}.$$

The expected payoff depends on x, the result of the vote, which itself depends on communication. There are three sources of uncertainty for x: the state of the world θ, the designated source of the message, and the success or failure of each fact-checking device. The following expression is a decomposition of $E(u_i(x, \theta)|(\boldsymbol{r}_i, \boldsymbol{r}_{-i}))$ using these different sources. I explain how the decomposition works in the subsequent paragraph, using the following notations:

- $Z_{S_i(j)}(\boldsymbol{r})$ is the number of vote for 1 in subset $S_i(j)$, when players play the strategy profile $\boldsymbol{r}$. This number is a random variable depending on how communication plays out. Particularly, it depends on the success or not of the fact-checking device, which itself depends on fact-checking strategies $\boldsymbol{r}$.

- $\mathbb{B}_i(j)$ is the event "the source is a biased agent in $S_i(j)$".

With this notation, for a fixed $\boldsymbol{r}_{-i}$ we get:

$$E(u_i(x,\theta)|(\boldsymbol{r}_i,\boldsymbol{r}_{-i})) = \mu_0 \times 0 + (1-\mu_0) \times \Bigg(\frac{1}{N}\frac{B}{N} + \tag{2}$$
$$\sum_{j\in N_i}\Bigg[\frac{U_i(j)}{N}\times\frac{B}{N} + \frac{B_i(j)}{N}\Big[\frac{1+E(Z_{S_i(j)}(\boldsymbol{r}_{-i})|\theta=0\cap\mathbb{B}_i(j))}{N}$$
$$+\sum_{j'\in N_i\setminus\{j\}}\Big[r_{ij'}(\frac{B_i(j')}{N}) + (1-r_{ij'})\frac{E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta=0\cap z_i=0\cap m_{ji}=1)}{N}\Big]\Big]\Bigg]\Bigg).$$

1. "$\mu_0 \times 0$": if $\theta = 1$, then every source (unbiased and biased) will create a trustful message equal to one. Every messenger will transmit, and because the message is true, fact-checking devices will never block the message (this is true independently of fact-checking strategies, because of the no false-positive assumption). Therefore, all agents will get a message equal to 1 and therefore vote for 1 (remember that agents play in Full Communication in the second stage). We get $(x-\theta)^2 = 0$.

2. "$(1-\mu_0)\times\ldots$": if $\theta = 0$, then biased agents will lie if they are the source, and the spread of this lie will depend on fact-checking strategies. To assert x, I, therefore, decompose the event again, depending on where the source is. The source can be on any branch linked to i (or i itself). Each term of the biggest sum is a different branch. If i is the source, then they create a message equal to 0, and only biased agents will vote for one (this happens with probability $\frac{1}{N}$). If i is not the source, I decompose further on each branch depending on if the source is biased or unbiased:

   - "$\frac{U_i(j)}{N}\times\frac{B}{N}$": if the source is unbiased, then the message created will be 0. In this case, all unbiased agents vote 0 (they either receive a message equal to 0 or the message is blocked, and they vote according to their priors, which are favorable to 0) and all biased agents vote 1. Hence, $P(x=1) = \frac{B}{N}$.
   - "$\frac{B_i(j)}{N}\Big[\ldots$": if the source is biased, the message created will be equal to 1. Unbiased agents who receive this message will misvote and lower $E[(x-\theta)^2]$. The extend of such agents depends on the spread of misinformation, which depends on fact-checking strategies. I isolate branches on which i can have an impact by fact-checking and branches on which they cannot.

– When agents play profile $\boldsymbol{r}$, $\frac{E(Z_{S_i(j)}(\boldsymbol{r})|\theta=0\cap \mathbb{B}_i(j))}{N}$ is the expected number of votes for 1 inside of $S_i(j)$ knowing that the source is a biased agent in $S_i(j)$ and $\theta = 0$. Note that i cannot have an impact on this number: messages from a source in $S_i(j)$ do not go through i to attain anyone in $S_i(j)$, $Z_{S_i(j)}$ only depends on the strategy of other players. Hence, for a fixed $\boldsymbol{r}_{-i}$, we can write with a slight abuse of notation $E(Z_{S_i(j)}(\boldsymbol{r})|\theta = 0 \cap \mathbb{B}_i(j)) = E(Z_{S_i(j)}(\boldsymbol{r}_{-i})|\theta = 0 \cap \mathbb{B}_i(j))$. The additional 1 is i themselves: it is assumed that i cannot see the result of their own fact-checking device.

– If the source is a biased agent in $S_i(j)$, i can impact anyone in $\cup_{j' \in N_i} S_{j'}(i)$ with her fact-checking strategies. For each fact-checked edge ij', $r_{ij'}$ is the probability the message is blocked by the fact-checking device. If this occurs, only biased agents in $S_{j'}(i)$ ($B_{j'}(i)$) will vote for 1. If not, then when agents play $\boldsymbol{r}$, $E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta = 0 \cap z_i = 0 \cap m_j i = 1))$ captures the expected number of votes for 1 once the message has crossed ij'. Note that this is independent of $r_{ij'}$. Hence for any fixed $\boldsymbol{r}_{-i}$, we can write with a slight abuse of notation $E(Z_{S_i(j')}(\boldsymbol{r})|\theta = 0 \cap z_i = 0 \cap m_j i = 1)) = E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta = 0 \cap z_i = 0 \cap m_j i = 1))$.

- "$\Pi_{j \in N_i}(1 - r_{ij})\frac{1}{N}$" represents the vote of i themselve. I assume that i (wrongly) votes for 1 when

Step 2: Show that the maximum of $U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i})$ is obtain for $r_{ij} \in \{0, 1\}, \forall j \in N_i$.

By inspection, $U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i})$ is linear in $r_i j$ for all $j \in N_i$. This implies that the maximum $U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i})$ is obtain for $r_{ij} \in \{0, 1\}, \quad \forall j \in N_i$.

□

With this useful lemma, we can move to the prove of 4.

Proof. Lemma 1 states that unbiased agents play $r = 0$ or $r = 1$ in dominant strategies. For a given edge ij, the choice of $r_{ij} = 0$ or $r_{ij} = 1$ will be determine by the sign of

$$U_i((r_{i,j} = 0, \boldsymbol{r}_{-i})) - U_i((r_{i,j} = 1, \boldsymbol{r}_{-i})).$$

Using the formula for U_i in the proof of Lemma 1, we can express this difference like this:

$$\begin{aligned} &U_i((r_{i,j} = 0, \boldsymbol{r}_{-i})) - U_i((r_{i,j} = 1, \boldsymbol{r}_{-i})) \\ &\qquad = -(1 - \mu_0)\lambda_U \Big(\sum_{j' \in N_i \setminus \{j\}} \Big[\frac{E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta = 0 \cap z_i = 0 \cap m_j i = 1)}{N} - \frac{B_i(j')}{N} \Big] \Big) + c \times 1. \end{aligned}$$

Note that:

$$E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta = 0 \cap z_i = 0 \cap m_j i = 1) \leq |S_i(j')|.$$

Indeed, the maximum number of people who can vote for 1 in $S_i(j')$ is the number of people in that subset: $|S_i(j')|$.

This means that:

$$-\lambda_U(\sum_{j'\in N_i\setminus\{j\}}\frac{|S_i(j')|-|B_i(j')|}{N}))+c=-\lambda_U(\sum_{j'\in N_i\setminus\{j\}}\frac{|U_i(j')|}{N}))+c\leq 0,$$

is a sufficient condition for $U_i((r_{i,j}=0,\boldsymbol{r}_{-i}))\leq U_i((r_{i,j}=1,\boldsymbol{r}_{-i}))$.
Rewritting, $U_i((r_{i,j}=0,\boldsymbol{r}_{-i}))\geq U_i((r_{i,j}=1,\boldsymbol{r}_{-i}))$ if

$$c\geq\lambda_U\frac{|U_i(j)|-1}{N}.$$

If this condition is satisfied then agent i plays 0 in dominant strategy.

Finally, we write $\underline{c}=\lambda_U\max_{i\in\mathcal{U}}(\max_{j\in N_i}(\frac{|U_i(j)|-1}{N})$. If $c\geq\underline{c}$, then all unbiased agents will chose $r=0$ on all of their edges.

□

Proof of Lemma 3 To prove lemma 3, I first need to prove lemma 2 that identifies dominant strategies for unbiased agents.

Lemma 2. *Take j a biased agent and and denote $\boldsymbol{r}_{-j}$ other agents' fact-checking profile. Assume that this profile is such that biased agents play at least their minimum trust inducing strategy given the profile of unbiased agents (see Proposition 2). Assume a Full Communication equilibrium is played in the second stage. Then, for all $i\in\mathcal{N}_j$ and all $r'_{ji}\in[0,1]$ we have either:*

$$U_j(r_{ji}=0,\boldsymbol{r}_{-j})\geq U_j(r'_{ji},\boldsymbol{r}_{-j})$$

or

$$U_j(\boldsymbol{r}^t_{ij}(\boldsymbol{r}_{\mathcal{U}}),\boldsymbol{r}_{-j})\geq U_j(r'_{ji},\boldsymbol{r}_{-j}),$$

where $r^t_{ji}(\boldsymbol{r}_{\mathcal{U}})$ is their minimum trust inducing strategy given $\boldsymbol{r}_{\mathcal{U}}$, the fact-checking strategies of unbiased agents in $\boldsymbol{r}_{-j}$.

Proof. In this proof, I first express the expected utility of any unbiased agent i as a function of their fact-checking behavior. This function is linear in i's strategies, implying that the maximum excepted utility is obtained either playing 0 or 1.

Step 1: Express unbiased agents payoff as a function of their fact-checking behavior.

Take i an unbiased agent and fix the profile of other players fact-checking strategies, that is $\boldsymbol{r}_{-i}$. Suppose that all agents (including i) play Full Communication strategies in the second stage.

I want to express the ex-ante utility of i as a function of their fact-checking strategies, that is $\boldsymbol{r}_i$, the vector of fact-checking from i on all edges they are connected to: $\boldsymbol{r}_i = (r_{ij})_{j \in N_i}$. I write this function as $U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i})$, where $\boldsymbol{r}_{-i}$ is fixed:

$$U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i}) = E(u_i(x,\theta)|(\boldsymbol{r}_i,\boldsymbol{r}_{-i})) = -\lambda_U E((x-\theta)^2|(\boldsymbol{r}_i,\boldsymbol{r}_{-i})) - c\sum_{j\in N_i} r_{ij}.$$

The expected payoff depends on x, the result of the vote, which itself depends on communication. There are three sources of uncertainty for x: the state of the world θ, the designated source of the message, and the success or failure of each fact-checking device. The following expression is a decomposition of $E(u_i(x,\theta)|(\boldsymbol{r}_i,\boldsymbol{r}_{-i}))$ using these different sources. I explain how the decomposition works in the subsequent paragraph, using the following notations:

- $Z_{S_i(j)}(\boldsymbol{r})$ is the number of vote for 1 in subset $S_i(j)$, when players play the strategy profile $\boldsymbol{r}$. This number is a random variable depending on how communication plays out. Particularly, it depends on the success or not of the fact-checking device, which itself depends on fact-checking strategies $\boldsymbol{r}$.
- $\mathbb{B}_i(j)$ is the event "the source is a biased agent in $S_i(j)$".

With this notation, for a fixed $\boldsymbol{r}_{-i}$ we get:

$$\begin{aligned} E(u_i(x,\theta)|(\boldsymbol{r}_i,\boldsymbol{r}_{-i})) = \mu_0 \times 0 + (1-\mu_0) \times \Bigg(& \frac{1}{N}\frac{B}{N} + \\ & \sum_{j\in N_i} \Bigg[\frac{U_i(j)}{N} \times \frac{B}{N} + \frac{B_i(j)}{N} \Bigg[\frac{1 + E(Z_{S_i(j)}(\boldsymbol{r}_{-i})|\theta = 0 \cap \mathbb{B}_i(j))}{N} \\ & + \sum_{j'\in N_i\setminus\{j\}} \Big[r_{ij'}(\frac{B_i(j')}{N}) + (1-r_{ij'})\frac{E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta = 0 \cap z_i = 0 \cap m_{ji} = 1)}{N} \Big] \Bigg] \Bigg] \Bigg). \end{aligned} \tag{3}$$

1. "$\mu_0 \times 0$": if $\theta = 1$, then every source (unbiased and biased) will create a trustful message equal to one. Every messenger will transmit, and because the message is true, fact-checking devices will never block the message (this is true independently of fact-checking strategies, because of the no false-positive assumption). Therefore, all agents will get a message equal to 1 and therefore vote for 1 (remember that agents play in Full Communication in the second stage). We get $(x-\theta)^2 = 0$.
2. "$(1-\mu_0) \times \ldots$": if $\theta = 0$, then biased agents will lie if they are the source, and the spread of this lie will depend on fact-checking strategies. To assert x, I, therefore, decompose the event again, depending on where the source is. The source can be on

any branch linked to i (or i itself). Each term of the biggest sum is a different branch. If i is the source, then they create a message equal to 0, and only biased agents will vote for one (this happens with probability $\frac{1}{N}$). If i is not the source, I decompose further on each branch depending on if the source is biased or unbiased:

- "$\frac{U_i(j)}{N} \times \frac{B}{N}$": if the source is unbiased, then the message created will be 0. In this case, all unbiased agents vote 0 (they either receive a message equal to 0 or the message is blocked, and they vote according to their priors, which are favorable to 0) and all biased agents vote 1. Hence, $P(x = 1) = \frac{B}{N}$.
- "$\frac{B_i(j)}{N}\Big[\ldots$": if the source is biased, the message created will be equal to 1. Unbiased agents who receive this message will misvote and lower $E[(x-\theta)^2]$. The extend of such agents depends on the spread of misinformation, which depends on fact-checking strategies. I isolate branches on which i can have an impact by fact-checking and branches on which they cannot.
  - When agents play profile $\boldsymbol{r}$, $\frac{E(Z_{S_i(j)}(\boldsymbol{r})|\theta=0\cap \mathbb{B}_i(j))}{N}$ is the expected number of votes for 1 inside of $S_i(j)$ knowing that the source is a biased agent in $S_i(j)$ and $\theta = 0$. Note that i cannot have an impact on this number: messages from a source in $S_i(j)$ do not go through i to attain anyone in $S_i(j)$, $Z_{S_i(j)}$ only depends on the strategy of other players. Hence, for a fixed $\boldsymbol{r}_{-i}$, we can write with a slight abuse of notation $E(Z_{S_i(j)}(\boldsymbol{r})|\theta = 0 \cap \mathbb{B}_i(j)) = E(Z_{S_i(j)}(\boldsymbol{r}_{-i})|\theta = 0 \cap \mathbb{B}_i(j))$. The additional 1 is i themselves: it is assumed that i cannot see the result of their own fact-checking device.
  - If the source is a biased agent in $S_i(j)$, i can impact anyone in $\cup_{j' \in N_i} S_{j'}(i)$ with her fact-checking strategies. For each fact-checked edge ij', $r_{ij'}$ is the probability the message is blocked by the fact-checking device. If this occurs, only biased agents in $S_{j'}(i)$ ($B_{j'}(i)$) will vote for 1. If not, then when agents play $\boldsymbol{r}$, $E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta = 0 \cap z_i = 0 \cap m_j i = 1))$ captures the expected number of votes for 1 once the message has crossed ij'. Note that this is independent of $r_{ij'}$. Hence for any fixed $\boldsymbol{r}_{-i}$, we can write with a slight abuse of notation $E(Z_{S_i(j')}(\boldsymbol{r})|\theta = 0 \cap z_i = 0 \cap m_j i = 1)) = E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta = 0 \cap z_i = 0 \cap m_j i = 1))$.
- "$\Pi_{j \in N_i}(1 - r_{ij})\frac{1}{N}$" represents the vote of i themselve. I assume that i (wrongly) votes for 1 when

Step 2: Show that the maximum of $U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i})$ is obtain for $r_{ij} \in \{0,1\}, \forall j \in N_i$.

By inspection, $U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i})$ is linear in $r_i j$ for all $j \in N_i$. This implies that the maximum $U_i(\boldsymbol{r}_i|\boldsymbol{r}_{-i})$ is obtain for $r_{ij} \in \{0,1\}, \quad \forall j \in N_i$.

□

With this useful lemma, we can move to the prove of 3.

Proof. Suppose unbiased agents plays a fixed strategy-profile $\boldsymbol{r}_{\mathcal{U}}$. Lemma 2 states that, given this unbiased fact-checking strategy profile, biased agents will either not fact-check, or play

minimum trust inducing profile.

Consider j biased and their strategy r_{ji}. We want to show that j prefers $r^t_{ji}(\boldsymbol{r}_{\mathcal{U}}$ to $r_{ji} = 0$.

From the expressions of ex-ante utility in the proof of Lemma 2 we can write:

$$\begin{aligned}
&U_j(r_{ji} = r^t_{ji}(\boldsymbol{r}_{\mathcal{U}}, \boldsymbol{r}_{-j}) - U_j(r_{ji} = 0, \boldsymbol{r}_{ji}) = \\
&- c \times r^t_{ji}(\boldsymbol{r}_{\mathcal{U}} - \lambda_B(\mu_0 \sum_{i' \in N_j \setminus i} \frac{S'_i(j)}{N} \times (-1)\frac{U_j(i)}{N} \\
&+ (1-\mu_0) \sum_{i' \in N_j \setminus i} P(m^{in}_j(i')) = 1 | \mathbb{S}_{i'}(j) \cap \theta = 0) \\
&\times (\frac{(1 - r^t_{ji}(\boldsymbol{r}_{\mathcal{U}})(S_j(i') - E(Z_{S_j(i)}(\boldsymbol{r}_{-j})|\theta = 0 \cap \mathbb{S}_i(j)) - U_j(i)}{N}).
\end{aligned}$$

Suppose that minimum truth inducing is played by every other biased agents. Note that, from proposition 3 , $E(Z_{S_j(i')}(\boldsymbol{r}_{-j})|\theta = 0 \cap \mathbb{B}_j(i))$, $P(m^{in}(i')) = 1|\mathbb{S}_{i'}(j) \cap \theta = 0)$ and $r^t_{ji}(\boldsymbol{r}_{\mathcal{U}}$ are uniquely determined by the network structure and $\boldsymbol{r}_{\mathcal{U}}$[6]. Hence, we have:

$$U_j(r_{ij} = r^t_{ij}(\boldsymbol{r}_{\mathcal{U}}) \geq U_j(r_{ji} = 0) \iff c \leq \lambda_B g_{ji}(G, \boldsymbol{r}_{\mathcal{U}}),$$

where $g_{ji}(.)$ is some function of the network structure G and unbiased fact-checking profile $\boldsymbol{r}_{\mathcal{U}}$.

Let $\bar{c}_{\boldsymbol{r}_{\mathcal{U}}} = \lambda_B \max_{j \in \mathcal{B}}(\max_{i \in \mathcal{N}_j}(g_{ji}(G, \boldsymbol{r}_{\mathcal{U}}))$.

If $c \leq \bar{c}_{\boldsymbol{r}_{\mathcal{U}}}$, then no *biased* agents have incentive to deviate from minimum trust inducing strategy.

□

A.3. Full Communication equilibria

Proof of Theorem 2

Proof. I show that the fact-checking profile $(\mathbf{1}_{\mathcal{U}}, \boldsymbol{r}^t_{\mathcal{B}}(\mathbf{1}_{\mathcal{U}}))$, where $\mathbf{1}_{\mathcal{U}}$ are unbiased fact-checking strategies where they all fact-check and $\boldsymbol{r}^t_{\mathcal{B}}(\mathbf{1}_{\mathcal{U}})$ is biased minimum trust-inducing strategies induced by the unbiased profile, holds as a PBE.

By backward induction, I first show that there is no incentive to deviate in the second stage if we assume that equilibrium fact-checking strategies are played in the second stage. Then, I show that given second stage play, there is no incentive to deviate in the first stage.

Step 1: Deviation in the second stage
Consider the fact-checking strategy profile $(\mathbf{1}_{\mathcal{U}}, \boldsymbol{r}^t_{\mathcal{B}}(\mathbf{1}_{\mathcal{U}}))$. According to Proposition 2, such a profile is Full Communication compatible. By definition, this implies that agents do not

[6]Because we assume minimum trust inducing strategies, $\boldsymbol{r}_{-j}$ is pinned down by $\boldsymbol{r}_{\mathcal{U}}$

have the incentive to deviate for Full Communication equilibrium in the second stage.

Step 2: Deviation in the first stage

Unbiased agents. Assume agents play the fact-checking strategies $(\mathbf{1}_{\mathcal{U}}, \boldsymbol{r}^*_{\mathcal{B}})$ and Full Communication strategies in the second stage. Consider any unbiased agent i. From Lemma 2, we know that unbiased agents will play either 0 or 1 depending on the sign of

$$U_i((r_{i,j}=0,\boldsymbol{r}_{-i})) - U_i((r_{i,j}=1,\boldsymbol{r}_{-i})).$$

With $c=0$, we have:

$$\begin{aligned} U_i((r_{i,j}=0,&\boldsymbol{r}_{-i})) - U_i((r_{i,j}=1,\boldsymbol{r}_{-i})) \\ &= -(1-\mu_0)\lambda_U\Big(\sum_{j'\in N_i\setminus\{j\}}\Big[\frac{E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta=0\cap m_j i=1)}{N} - \frac{B_i(j')}{N}\Big]\Big). \end{aligned}$$

But for all $j'\in N_i$, we have:

$$E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta=0\cap m_j i=1) \geq B_i(j').$$

Indeed, $E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta=0\cap m_j i=1)$ is the expected number of agents voting for 1 when i transmitted a false message. Because we assume Full Communication strategies, biased agents will always vote for 1 whatever their beliefs are. Hence, we necessarily have $E(Z_{S_i(j')}(\boldsymbol{r}_{-i})|\theta=0\cap m_j i=1) \geq B_i(j')$. This implies:

$$U_i((r_{i,j}=0,\boldsymbol{r}_{-i})) \geq U_i((r_{i,j}=1,\boldsymbol{r}_{-i})).$$

Note that this inequality will often be strict because it is likely that some unbiased agents will believe that message spread by i when she does not fact-check it and vote for 1. Indeed, it is assumed throughout the paper that agents cannot see the outcome of their fact-checking device. An unbiased agent receiving a message equal to 1 (and believing it) will vote for 1 even if their fact-checking device intercepts the message. Unbiased agents likely believe such messages if all unbiased agents perfectly fact-check.

This implies that i cannot increase their payoff by deviating from perfectly fact-checking.

Biased agents. Suppose $c=0$, that agents play equilibrium strategies and consider a biased agent j. Because $c=0$, Proposition 3 applies, and j best-response to other agents strategies is their minimum trust-inducing strategy. Hence, j as no incentive to deviate either.

□

Proof of Theorem 4

Proof. In this proof, I first define a restricted game where agents are only allowed to play Full Communication compatible (FCC) strategies. I show that this game's strategy space is a closed non-empty convex space, and that payoff functions are quasiconcave. Kakutani's theorem implies that an equilibrium of this restricted game exists for any cost of c. In the second part, I show that under the theorem conditions, agents have no incentive to deviate from the restricted game's equilibrium strategies even if we allow them to play on the whole strategy space (*i.e.* if we lift restrictions).

Existence of equilibrium in the FCC-restricted game.

Suppose an instance of the game with graph G, preferences λ_B and λ_U and cost c. Let $\bar{R}$ be the set of Full Communication Compatible actions:

$$\bar{R} = \{r_{ij} \in [0,1], \forall ij \in E | r_{ij} \geq r^t_{ij}(\boldsymbol{r}_{\mathcal{U}}) \text{ if } i \in \mathcal{B}\}$$

If fact-checking actions are played in $\bar{R}$, then Full Communication strategies form an equilibrium of the second-stage for any play of the first-stage (any fact-checking). We can redefine the game as a simultaneous complete information game if we assume that agents play this equilibrium in the second stage. In such a game, the payoffs from a strategy profile are the expected payoff agents would get if this strategy profile was played in the first stage, and full communication strategies were played in the second stage.

We write $BR : (\boldsymbol{r}_{\mathcal{B}}, \boldsymbol{r}_{\mathcal{U}}) \to (\boldsymbol{r}_{\mathcal{B}}, \boldsymbol{r}_{\mathcal{U}})$ the best-response correspondence of this restricted game.

I use Kakutani's theorem for continuous games to show that this correspondence has a fixed-point (see, for example, Theorem 1.2. in Fudenberg & Tirole (1991)). To apply this result, I need to prove that the action space is non-empty, convex, and compact and that payoffs are continuous and quasiconcave in actions.

- Consider any fact-checking profile $\boldsymbol{r}_{\mathcal{B}} = \mathbf{1}$ where all biased agents fact-check perfectly. Because minimum trust-inducing strategy always exist in $[0,1]$ (from 2), we necessarily have $(\boldsymbol{r}_{\mathcal{U}}, \mathbf{1}) \in \bar{R}$ for any $\boldsymbol{r}_{\mathcal{U}}$. Hence $\bar{R}$ is non empty.
- **Convexity:** Lemma 3 bellow shows that $\bar{R}$ is convex.
- **Compacity.** As a closed and bounded subset of $\mathbb{R}^n$, $\bar{R}$ is compact.
- On $\bar{R}$, the payoff functions are continuous and linear, hence quasiconcave.

Applying Kakutani's theorem, we get that there is at least one equilibrium of the restricted game, that we write $(\boldsymbol{r}^*_{\mathcal{U}}, \boldsymbol{r}^*_{\mathcal{B}})$. Observe that an equilibrium exist for any instance of the game, and in particular, for any $c \geq 0$.

Equilibrium in the non-restricted game. In this second part, we fix the strategies $(\boldsymbol{r}^*_{\mathcal{U}}, \boldsymbol{r}^*_{\mathcal{B}})$ obtained in the previous step, and consider how agents best respond to these strategies if we allowed them to play on the whole strategy space. I will show that, if the theorem conditions are respected, then no agents have incentive to deviate from $(\boldsymbol{r}^*_{\mathcal{U}}, \boldsymbol{r}^*_{\mathcal{B}})$.

Step 1: Unbiased agents deviation

Let us first consider unbiased agents. Observe that even in the restricted game, unbiased agents can play on their full scope of action $[0, 1]$. Therefore, unbiased agents' best-response will not change if we move to the unrestricted game. This is true for any c.

Step 2: Biased agents deviation

Let us now consider the biased agents' best response. Remember that from Proposition 4, we know that unbiased agents either play 0 or their minimum trust inducing strategy (which can be 0). In a Full Communication Equilibrium, biased agents play their minimum trust inducing strategy. We therefore only need to check that biased agents are not willing to deviate towards 0. In the following when I write "biased agents are willing to fact-check" I mean that they play their minimum trust inducing strategy.

According to Proposition 4, for all $c \leq \bar{c}_{\boldsymbol{r}^*_{\mathcal{U}}}$, biased agents best respond to $\boldsymbol{r}^*_{\mathcal{U}}$ by their minimum trust inducing strategies, *i.e.* $\boldsymbol{r}^t_{\mathcal{B}}(\boldsymbol{r}^*_{\mathcal{U}}) \in BR(\boldsymbol{r}^*_{\mathcal{B}})$.

Using this fact, I will expose several conditions under which the equilibrium holds. I will proceed in the following way:

1. I show that there is a maximum cost, c_M, under which biased agents are willing to fact-check *whatever is the fact-checking behavior of unbiased agents.*
2. I show that there is a minimum cost, $\underline{c}$, above which *unbiased* agents will always choose a fact-checking of 0.
3. I present a condition such that $\underline{c} \leq c_M$.
4. I conclude that when this condition is satisfied there is a maximum cost $\bar{c} \geq c_M$ such that biased agents are willing to fact-check for all cost smaller than c_0, where c_0 is the maximum cost biased agents are ready to pay when they know that unbiased agents do not fact-check.

Step 2.1: Condition 1 - maximum cost for fact-checking in dominant strategy.

If $c \leq \min_{\boldsymbol{r}_{\mathcal{U}}}(\bar{c}_{\boldsymbol{r}_{\mathcal{U}}})$, then we must have $c \leq \bar{c}_{\boldsymbol{r}^*_{\mathcal{U}}}$, hence biased agents do not have incentive to deviate. Writing $c_M = \min_{\boldsymbol{r}_{\mathcal{U}}}(\bar{c}_{\boldsymbol{r}_{\mathcal{U}}})$ and taking step 1 into account, we can sum this up with the following condition:
If $c \leq c_M$ then $(\boldsymbol{r}^*_{\mathcal{U}}, \boldsymbol{r}^*_{\mathcal{B}})$ holds as an equilibrium in the unrestricted game.

Step 2.2 Minimum cost for no unbiased fact-checking.

Proposition 2 (on best response of unbiased agents), states that there exist a cost $\underline{c} = \lambda_U \max_{i\in\mathcal{U}}(\max_{j\in N_i}(\frac{|U_i(j)|-1}{N})$, such that if $c \geq \underline{c}$, no unbiased agents fact-check (in dominant strategy).

Step 2.3 Condition 3 - Ordering of $\underline{c}$ and $\bar{c}$.

We want to find a condition such that $\underline{c} \leq c_M$. If this inequality is true, then it means that, as we increase cost, unbiased agents will stop fact-checking *before* biased agents stops fact-checking in dominant strategy. Using the expression for $\underline{c}$ and c_M we can write:

$$\begin{aligned}\underline{c} \leq c_M &\iff \lambda_U \max_{i\in\mathcal{U}}(\max_{j\in N_i}(\frac{|U_i(j)|-1}{N}) \leq \min(\lambda_B \max_{j\in\mathcal{B}}(\max_{i\in\mathcal{N}_j}(g_{ij}(G,\boldsymbol{r}_{\mathcal{U}})))\\ &\iff \frac{\lambda_B}{\lambda_U} \geq \frac{\max_{i\in\mathcal{U}}(\max_{j\in N_i}(\frac{|U_i(j)|-1}{N})}{\min_{\boldsymbol{r}_{\mathcal{U}}}(\max_{j\in\mathcal{B}}(\max_{i\in\mathcal{N}_j}(g_{ij}(G,\boldsymbol{r}_{\mathcal{U}})))}.\end{aligned}$$

For simplicity, let's write $\tilde{g} = \frac{\max_{i\in\mathcal{U}}(\max_{j\in N_i}(\frac{|U_i(j)|-1}{N})}{\min_{\boldsymbol{r}_{\mathcal{U}}}(\max_{j\in\mathcal{B}}(\max_{i\in\mathcal{N}_j}(g_{ij}(G,\boldsymbol{r}_{\mathcal{U}})))}$. The nominator only depends on the number of unbiased agents in each branch, hence it only depends on the network structure. The denominator depends on $g_{ij}(.)$ who depends on network structure and $\boldsymbol{r}_{\mathcal{U}}$. But since we choose the value $\boldsymbol{r}_{\mathcal{U}}$ that minimize it, it only depends on the network structure. Therefore, for a fixed network, $\tilde{g}$ is a constant.

Conclusion of step 2.

From the previous step, there exists a $\bar{g}$, that only depends on network structure, such that $\underline{c} \leq c_M$ iff $\frac{\lambda_B}{\lambda_U} \geq \bar{g}$. Going back to the definitions of $\underline{c}$ and c_M, this means that unbiased agents stop fact-checking before biased agents stop fact-checking in dominant strategy. We can look at what happens on each side of $\underline{c}$ now:

- For any $c \leq \underline{c}$, we also have $c \geq c_M$. Therefore, we can be sure that biased agents will play fact-check in dominant strategy. Said otherwise, they best respond to $\boldsymbol{r}^*_{\mathcal{U}}$ by their minimum trust inducing strategies.
- For $c \geq \underline{c}$, we might be worried that $c \geq c_M$ and that biased agents do not want to fact-check in dominant strategy. We know that *unbiased* agents play 0 in the dominant strategy. Hence for any equilibrium we found in the restricted game, we have $\boldsymbol{r}^*_{\mathcal{U}} = 0$ (as a vector). Going back to the beginning of step 1, we said that from Proposition 3, for all $c \leq \bar{c}_{\boldsymbol{r}^*_{\mathcal{U}}}$, biased agents will best respond to $\boldsymbol{r}^*_{\mathcal{U}}$ by their minimum trust inducing strategies. Hence, if c is such that $\underline{c} \leq c \leq c_{\mathbf{0}}$ (where $c_{\mathbf{0}}$ is the maximal cost defined in Proposition 2 when we assume that unbiased agents fact-check 0), then biased agents best respond with their minimum trust inducing strategy.

Hence, if $\frac{\lambda_B}{\lambda_U} \geq \bar{g}$ and for any $c < c_{\mathbf{0}}$, biased agents have no incentive to deviate from the restricted game equilibrium even if allowed to play in the unrestricted strategy space.

General conclusion.

In the first part of this proof, I have shown that if we restrict the strategy space to Full Communication Compatible strategies, then an equilibrium exists. In the second part, I start from this point (equilibrium of restricted game) and relax restrictions on the strategy space. I first remind that unbiased agents do not have an incentive to deviate from this point because their strategies were not restricted in the first place. Then I show if the following conditions are met:

1. $\frac{\lambda_B}{\lambda_U} \geq \bar{g}$ and,
2. $c \leq c_{\mathbf{0}}$,

then biased agents do not have incentives to deviate even if we allow them to play on the whole strategy space.

□

Lemma 3. *The space of Full Communication Compatible strategies, defined as:*

$$\bar{R} = \{r_{ij} \in [0,1], \forall ij \in E | r_{ij} \geq r^t_{ij}(\boldsymbol{r}_{\mathcal{U}}) \text{ if } i \in \mathcal{B}\}$$

is convex.

Proof. To show that $\bar{R} = \{r_{ij} \in [0,1], \forall ij \in E | r_{ij} \geq r^t_{ij}(\boldsymbol{r}_{\mathcal{U}}) \text{ if } i \in \mathcal{B}\}$ is convex, I need to show that the function r^t_{ji} is convex for all ji such that j is biased.

Step 1: *Formulating the problem.*

Saying that $\boldsymbol{r}^t_{ji}$ is convex is equivalent to say that for any $\boldsymbol{r'}_{\mathcal{U}}, \boldsymbol{r''}_{\mathcal{U}}$, and any $\lambda \in [0,1]$:

$$r^t_{ji}(\lambda \boldsymbol{r'}_{\mathcal{U}} + (1-\lambda)\boldsymbol{r''}_{\mathcal{U}}) \leq \lambda r^t_{ji}(\boldsymbol{r'}_{\mathcal{U}}) + (1-\lambda) r^t_{ji}(\boldsymbol{r''}_{\mathcal{U}}).$$

To simplify notation let's rewrite that as:

$$r^t_{ji}(\boldsymbol{r}^{mix}_{\mathcal{U}}) \leq \lambda r^t_{ji}(\boldsymbol{r'}_{\mathcal{U}}) + (1-\lambda) r^t_{ji}(\boldsymbol{r''}_{\mathcal{U}}). \tag{4}$$

From Proposition 2:

$$r^t_{ji}(\boldsymbol{r}_{\mathcal{U}}) = 1 - \frac{\mu_0}{(1-\mu_0) P_{\boldsymbol{r}_{\mathcal{U}}}(m^{out}_i = 1 | \theta = 0)},$$

where $P_{\boldsymbol{r}_{\mathcal{U}}}(m_j = 1 | \theta = 0)$ is the probability that j sends a message equal to one to i when the state is 0, the source of the message is in $S_j(i)$ and all other agents (other than j) play profile $(\boldsymbol{r}_{\mathcal{U}}, \boldsymbol{r}^t_{\mathcal{B}}(\boldsymbol{r}_{\mathcal{U}}))$. This quantity depends on the probability that j himself receives a message, which depends on fact-checking strategies upstream of ji.

Step 2: *Simplifying the problem.*

To simplify notations, let's define the following function:

$$p_{ij} : \begin{cases} [0,1]^{|\mathcal{E}_{\mathcal{U}}|} \to [0,1] \\ \boldsymbol{r}_{\mathcal{U}} \to P_{\boldsymbol{r}_{\mathcal{U}}}(m_j = 1 | \theta = 0). \end{cases}$$

To determine the convexity of r^t_{ji}, we need to characterize some properties of p_{ij}.

Using the definition of $r^t_{ji}(\boldsymbol{r}_{\mathcal{U}})$ and p_{ij}, we can find a simpler formulation for (1). We get that r^t_{ji} is convex only and only if:

$$\frac{1}{p_{ij}(\boldsymbol{r}^{mix}_{\mathcal{U}})} \geq \frac{\lambda}{p_{ij}(\boldsymbol{r'_{\mathcal{U}}})} + \frac{1-\lambda}{p_{ij}(\boldsymbol{r''_{\mathcal{U}}})},$$

which we can rewrite as:

$$p_{ij}(\boldsymbol{r}^{mix}_{\mathcal{U}}) \leq \frac{p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) \times p_{ij}(\boldsymbol{r}''_{\mathcal{U}})}{\lambda p_{ij}(\boldsymbol{r}''_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}'_{\mathcal{U}})}.$$

This condition is hard to prove in general. We can make the observation that

$$\frac{p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) \times p_{ij}(\boldsymbol{r}''_{\mathcal{U}})}{\lambda p_{ij}(\boldsymbol{r}''_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}'_{\mathcal{U}})} \geq \lambda p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}''_{\mathcal{U}}).$$

Let's show why. Consider the sign of $\frac{p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) \times p_{ij}(\boldsymbol{r}''_{\mathcal{U}})}{\lambda p_{ij}(\boldsymbol{r}''_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}'_{\mathcal{U}})} - \lambda p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}''_{\mathcal{U}})$. Multiplying by the first term denominator we can get the following simplification:

$$\begin{aligned} p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) \times p_{ij}(\boldsymbol{r}''_{\mathcal{U}}) - & [\lambda p_{ij}(\boldsymbol{r}''_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}'_{\mathcal{U}})][\lambda p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}''_{\mathcal{U}})] \\ & = p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) p_{ij}(\boldsymbol{r}''_{\mathcal{U}}) 2\lambda(1-\lambda) + p_{ij}(\boldsymbol{r}'_{\mathcal{U}})^2 \lambda(1-\lambda) + p_{ij}(\boldsymbol{r}''_{\mathcal{U}})^2 \lambda(1-\lambda) \\ & = \lambda(1-\lambda)(p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) + p_{ij}(\boldsymbol{r}''_{\mathcal{U}}))^2. \end{aligned}$$

Because $\lambda \in [0,1]$ and $(p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) + p_{ij}(\boldsymbol{r}''_{\mathcal{U}}))^2 \geq 0$, it is clear that the expression is positive. Hence,

$$\frac{p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) \times p_{ij}(\boldsymbol{r}''_{\mathcal{U}})}{\lambda p_{ij}(\boldsymbol{r}''_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}'_{\mathcal{U}})} \geq \lambda p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}''_{\mathcal{U}})$$

.

This means that

$$p_{ij}(\boldsymbol{r^{mix}}_{\mathcal{U}}) \leq \lambda p_{ij}(\boldsymbol{r'}_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r''}_{\mathcal{U}}),$$

is a sufficient condition for (1), that is for the convexity of r^t_{ji}. Indeed, if this condition is true, we would have:

$$p_{ij}(\boldsymbol{r}^{mix}_{\mathcal{U}}) \leq \lambda p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}''_{\mathcal{U}}) \leq \frac{p_{ij}(\boldsymbol{r}'_{\mathcal{U}}) \times p_{ij}(\boldsymbol{r}''_{\mathcal{U}})}{\lambda p_{ij}(\boldsymbol{r}''_{\mathcal{U}}) + (1-\lambda) p_{ij}(\boldsymbol{r}'_{\mathcal{U}})}.$$

Summing up what we did up until now, we get the following proposition:

If p_{ij} is a convex function for all ij, then $\bar{R}$ is a convex space.

The rest of the proof shows that p_{ij} is indeed a convex function for all ij.

***Step 3**: Show that p_{ij} is a convex function for all ij.*

We can express p_{ij} in the following manner:

$$\begin{aligned} p_{ji}(\boldsymbol{r}_{\mathcal{U}}) &= P_{\boldsymbol{r}_{\mathcal{U}}}(m_j^{out} = 1 | \theta = 0) \\ &= \sum_{i' \in \mathcal{B}_i(j)} \frac{1}{|S_i(j)|} \Pi_{rm \in i' \to j} (1 - r_{rm}^{mix}), \end{aligned}$$

where $i' \to j$ is the set of all the edges (denoted by the indices rm) in the path from i' to j. Because the network is a tree, this path is unique.

The reasoning behind this expression is the following. If we assume Full Communication strategy, and if j is biased, then j always sends a message equal to 1 if they get one. $P_{\boldsymbol{r}_{\mathcal{U}}}(m_j^{out} = 1 | \theta = 0)$ is therefore equal to the probability that he gets a message equal to 1. If the state is 0 (as we assume it is), such an event can occur if only if a biased agent is the *source*. For j to receive a message emitted by such a source, it must be that all fact-checking devices between i and j fail, which is given by the probability $\prod_{rm \in i' \to j} (1 - r_{rm}^{mix})$ ($\prod$ is the product operator). If we sum this overall potential biased source (weighting by the probability that they are indeed source), we get the expression above.

We want to show the convexity of this expression. It is easier to do with a slightly more general framework. Consider a function $f(x_1, \cdots, x_n) = x_1 \times x_2 \cdots x_n$. If we can show that this function is convex for $x_i \in [0, 1]$, then it means that the above expression is convex (the sum of a convex function is convex). To do so, I will compute the Hessian matrix and show that it is positive semi-definite.

The second order partial derivatives of $f(.)$ can be computed as:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = x_1 x_2 \cdots x_{i-1} x_{i+1} \cdots x_{j-1} x_{j+1} \cdots x_n = \frac{f}{x_i x_j}.$$

Therefore the elements of Hessian are $[H]_{ij} = \frac{f}{x_i x_j}$. For an arbitrary vector $v \in \mathbb{R}^n_{++}$ (strictly positive real numbers) we have

$$v^T H v = \sum_{mn \in E} \sum_{kl \in E} f \frac{v_i}{x_i} \frac{v_j}{x_j} (1 - \delta_{ij}) > 0,$$

where $\delta(i - j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

$v^T H v$ is a positive on $\mathbb{R}^n_{++}$, thus the Hessian is positive definite. This implies that the function is *strictly* convex in $\mathbb{R}^n_{++}$. In $\mathbb{R}^n_{+}$ (including 0), one or more x_i can be null, in which case the Hessian is positive semi-definite, which implies that $f(.)$ is weakly convex (and that

is enough for us!).

We have $(1 - r_{ij}^{mix}) \in \mathbb{R}_+^n$ (for $n = |E|$), and we know that the sum of convex functions is convex. Hence, we directly get that $\sum_{i' \in \mathcal{U}^{unchk}} \frac{1}{|S_i(j)|} \Pi_{\mu \in i' \to j} (1 - r_{i'j}^{mix})$ is convex.

Conclusion: I have showed that for any edge ij where i is biased, p_{ij} is convex. From step 2, this implies that for any two fact-checking profile $\boldsymbol{r'_{\mathcal{U}}}$ and $\boldsymbol{r''_{\mathcal{U}}}$, we have:

$$\frac{1}{p_{ij}(\boldsymbol{r}_{\mathcal{U}}^{mix})} \geq \frac{\lambda}{p_{ij}(\boldsymbol{r'_{\mathcal{U}}})} + \frac{1-\lambda}{p_{ij}(\boldsymbol{r''_{\mathcal{U}}})}.$$

Traducing this in term of minimum trust inducing strategies from Step 1, this gives us:

$$r_{ji}^t(\lambda \boldsymbol{r}'_{\mathcal{U}} + (1-\lambda)\boldsymbol{r}''_{\mathcal{U}}) \leq \lambda r_{ji}^t(\boldsymbol{r}'_{\mathcal{U}}) + (1-\lambda) r_{ji}^t(\boldsymbol{r}''_{\mathcal{U}}).$$

Plugging this into the definition of $\bar{R}$, we obtain that $\bar{R}$ is a convex space.

□

Proof of Proposition 3

Proof. We write:

$$g = \frac{\max_{j \in \mathcal{B}}(\max_{i \in \mathcal{N}_j}(g_{ij}(G, \boldsymbol{r}_{\mathcal{U}} = \mathbf{0}))}{\max_{i \in \mathcal{U}}(\max_{j \in N_i}(\frac{|U_i j|-1}{N})}.$$

Suppose $\frac{\lambda_B}{\lambda_U} \geq g$. If we denote $\bar{c} = \bar{c}_{\mathbf{0}}$ the maximum cost that induces biased agents to play minimum trust inducing strategies when unbiased agents don't fact check (see lemma 3), the condition implies that $\underline{c} < \bar{c}$.

Consider now a cost $c \in [\underline{c}, \bar{c}]$. From lemma 4, we know that unbiased agents do not fact-check in dominant strategies. Furthermore, because $c \leq \bar{c}$, we know that biased agents' best response to this unbiased fact-checking profile will be their minimum trust inducing strategies. Applying the formula for minimum trust inducing strategies when unbiased agents do not fact-check from 1 gives the result.

□