Fake news debunking in social networks: an evolutionary approach to behaviors

Gianluca Principini

Complex Systems and Network Science
Computer Science Department - Bologna University

Abstract. The online community feedbacks are important to understand public opinion changes and other group dynamics. Nowadays social networks represent the most important media of news distribution. Unfortunately, not everyone fact checks these news, contributing to the spreading of falseness, while other people do fact check in order to demistify misinformative contents. Interesting social behaviors emerge from interactions between these two categories of people in a dynamic social network in which each component tries to maximize its credibility to the eyes of others.

1 Introduction

Social media have changed the way informations spread among the population. Web 2.0 is offering new tools for the creation of new contents to everyone, including non professionals. As consequences is very easy to create fake news websites which resemble proper journalist work and release fake contents. The major problem of this fake-news spreading is the loss of trust in traditional and verified media. This can lead to the manipulation of the public opinion, with huge effects on the electorate. [2]. The presence of some cognitive biases, such as confirmation bias and implicit biases related to concordance of opinion among relationships, contribute to fake-news virality. To counter this phenomenon, internet debunkers emerged, users specialized in fact-checking of the news that circulates on the web. This one tries to expose the falseness of fake-news through rational reasoning and references to authoritative sources. A social network is a complex interconnected system in which each component is an active entity that can influence evolution of the entire system. As showed by Axelrod [1], Game Theory is an useful tool that could be used to understand how complex social behaviors emerge and evolve through time. Thus, the aim of the project is to apply game theory to social networks, in order to understand and highlight evolutionary mechanisms which control behavioral changes in a community as a consequence of online news spreading. The model will focus, in particular, on the creation of polarized subgroups, as result of interactions, and stem problems by simulating different approaches. In preliminaries the concept of fake-news will be explored, some cognitive biases governing their spreading will be explained deeper and then game theory will be introduced. In main part will be explained the model and some results of simulations will be showed.

2 Preliminaries

2.1 Fake news

The term "Fake News" has become popular after the Donald Trump's first press conference as president-elect in November 2016, and can refer to different types of topics like politics, science and conspiracy theory. A fake news could be defined as a news article, verifiably false, which can deceive population misleading readers using inaccuracies. This definition also includes articles produced by satirical websites that could be misunderstood as factual. Since the earliest days of the internet, fake news has circulated online in newsgroups made of conspiracy theorists cliques. In 2017 we receive news also in a variety of other online platforms like social media, in which almost everyone has potentially access to millions of followers. If an influencer goes down for a fake-news, there's a high probability that it will become viral, without a proper fact-checking. The results could be worse than expected when a fake-news story started its spreading, especially when they are written aiming to emotional sphere of people about controversial topics. A Stanford study evidenced how much fake news were involved in 2016 American elections and that some people still believe in several already debunked hoaxes largely spread in the last century [2].

2.2 Sociological and psychological background

But why do fake news proliferate so rapidly? There's a massive amount of research in social psychology about "naive realism" or "direct realism", which is the human tendency to believe that we see the world objectively, and that people who disagree must be uninformed, irrational or biased. This behavior leads to an exaggeration of differences between the self and the others, creating barriers to negotiation through several mechanisms:

1. *Bias blind spot*: the cognitive bias of recognizing cognitive and motivational biases in others while failing to recognize the same on the self. [4]
2. *False polarization*: the cognitive bias that involves interpreting others' views as more extreme than they really are. [5]
3. *Reactive devaluation*: that occurs when a proposal is devalued if it appears to originate from an antagonist, creating a barrier for social conflict resolution. [6]

It means that people who hold strong opinions on complex issues are likely to examine relevant empirical evidence in a biased manner, so when people hear or see something consistent with their beliefs, there is a tendency to believe it, while they are more critical of researches not aligned with their beliefs. [8] [9] The tendency to search for, interpret and recall information selectively, in a way that confirms preexisting hypotheses is a well known cognitive bias called *confirmation bias*. This leads to an attitude polarization, a phenomenon where people's attitudes or beliefs strengthen and become more extreme as they engage in intensive thought about the attitude object [10]. It has also been proved

that even when people admit the using of biased sources, they still think they reach an objective conclusion. [7]. The concept of attitude polarization could be extended in a group context. When people are placed in a group, this one has some overriding attitude towards the situation, so decision and opinions of people become more extreme than their initial inclination. This phenomenon has been observed in the newest researches focused on social media, it exists even when a group is not physically together. Owing to this technology, it is easier for people to curate their sources of information. Another factor could give rise to these behaviors: information filtering in social medias, which gives major priority to content shared by people who have same interests and values leads to the creation of filter bubbles, or digital echo chambers, in which users get trapped in a sort of monoculture, because these platforms obviously influence what a user sees and does not see, learning from what the user already sees. [11]
On the other hand, the figure of debunker emerged in social networks. A debunker is a person, or an organization, who tries to expose the sham or falseness of something, especially pseudoscientific researches, alternative medicine, conspiracy theories and other claimed paranormal phenomena. This term is also used in a more general sense, at attempts to discredit any opposing point of view. A failed debunking, however, can actually worsen misconception, with a backfire effect. This occurs when a debunker accidentally reinforce false beliefs by trying to correct them, so it is important to follow some guidelines while debunking. [12]
The understanding of behavioral changes and evolution of relationships, due to interactions among different people, through simulation, will probably make easy to comprehend better the rise of these phenomena and which countermeasures should be taken in order to limit them.

2.3 Game Theory

Game theory is designed to address situations in which the outcome of a person's decision depends on their choose and on the choices made by the people they are interacting with. Some contexts are literally games, others are not usually called games, but can be analyzed with the same tools. Game theory could be used to study situations in which decision makers interact with one another. The primary goal of game theory is to understand which *strategy*, meaning a complete set of rules that prescribes what a player should do in every possible situation for maximize his reward in a context, for a particular game and which behaviors tend to sustain themselves when carried out in a large population. [13] [14]
A game is any situation that could be represented with the following three aspects:

1. *Players*: a set of participants
2. *Strategy*: a set of options for how to behave
3. *Payoff*: received by each player depending on everyone's strategies

3 Main part

3.1 Formal definition of the game

In this section a game theory model in which several people connected each other in a realistically clustered social network interact, sharing or debunking fake news and misinformative contents to improve their payoff will be described. In order to reason about this game it is useful to make a few assumptions:

1. A player can only share, see and debunk fake news shared by other players he is connected to, those will be called *neighbors*.
2. A connection between two players has a *weight*, which indicates the confidence between them. This connection could either break up or not, like it happens in real social networks in which some friendships end.
3. Everything that a player cares about is summarized in its payoff, that would be, in this project, his credibility among the people he is connected to.
4. Each player knows his list of possible strategies, in this case the possibility of sharing a fake news or misinformative contents or not.
5. Players choose a new strategy depending on the reactions of their neighbors, once achieved the new credibility is less than the previous one in order to maximize it.

To understand better the structure of this game it is useful to define formally the ideas of social network and players:

Definition 1 (Social Network). *According to the previous assumptions, a social network could be defined as a weighted undirected graph as follows:*

$$S = (V, E)$$

With $V = \{v_1, v_2, ..., v_n\}$ *the set of players and* $E \subseteq \{\{v_i, v_j\} \mid v_i, v_j \in V \wedge i \neq j\}$ *the set of undirected weighted connections between players, the weight of* $\{v_i, v_j\} \in E$*, which in this game represents the confidence between* v_i *and* v_j*, will be referred to as* $w_{i,j} \in [0, 1]$*. If* $w_{i,j} \simeq 0$ *and the link removal option is enabled, the link is deleted, if* $w_{i,j} > 1$ *it will be set to 1.*

Definition 2 (Player). *A player* $v_i \in V$ *is a tuple:*

$$v_i = (\phi_i, \phi_i', C_i)$$

1. $\phi_i \in [0, 1]$ *the player's inclination to ignore, and debunk if the self-debunk option is enabled, the fake news he receives or debunk a fake news shared by one of his neighbors. Will be referred to as fact-checking attitude This could be seen as a probability that determines how a player i would probabilistically behave in each turn. For simplicity when it becomes greater than 1 it will be set to 1, when it become less than 0 it will be set to 0.*
2. $\phi_i' \in R$ *the fact-checking attitude changes. This value is set to zero at the beginning of each turn and represents the behavioral changes due to interactions with neighbors for each turn*

3. $C_i \in Z$ *the credibility of the players, that represents how much a player i is credible among his neighbors, this attribute will also be the payoff for the game.*

Definition 3 (Neighborhood). *The neighborhood of a player* $v_i \in V$ *is defined as:*

$$\delta(v_i) = \{v_j \in V \mid \{v_i, v_j\} \in E\}$$

Definition 4 (Degree). $\forall v_i \in V$*, the degree of* v_i *is defined as*

$$D(v_i) = |\delta(v_i)|$$

The average degree is $\bar{D} = \frac{1}{|V|} \sum_{v_i \in V} D(v_i)$

Indicating with N_{fc} the number of fact-checkers and N_{nfc} the number of non fact-checkers, the game starts with the following settings: $10 \leq N_{nfc}, N_{fc} \leq 150$ players scattered in a bidimensional space. In order to make the distribution of people and their relationships in the network more realistic, links are created aiming to create a spatially clustered network, according to an arbitrarily defined average degree. Let *maxdistance* the max possible distance between two players, so the diagonal of the bidimensional world, and $distance(v_i, v_j) \in [0, maxdistance]$ the distance function between two players in the space, if the relationship feedback is enabled the weights are initialized as it follows:

$$\forall \{v_i, v_j\} \in E : w_{i,j} = \frac{maxdistance - distance(v_i, v_j)}{maxdistance}$$

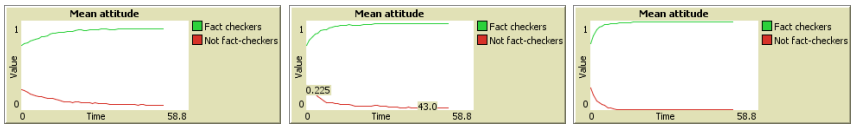
Otherwise it will be 0.5 for every relationship. That is, the more players are close together, the more they trust each other when the simulation starts. Other global parameters governing the simulation are: $\alpha \in [0, 0.5]$, the confirmation bias factor, described in preliminaries, $\beta \in [0, 0.1]$, the relationship modifier factor, useful to understand how the strength of confidence between people changes when they share the same behavior or interact with people who have different behaviors, taking into account what said in preliminaries, this modifier could also be disabled through a switch named relationship-feedback , $\gamma \in [0, 0.5]$, the effectiveness of debunking and $\epsilon \in [1, 5]$, the diffidence in debunks. However, during the experiments will be considered exclusively configuration classes which are more representative. At each turn, the spreading of a fake news works as showed by the FakeNewsPropagation algorithm **??**. At the end of each turn the new credibility for each player $v_i \in V$ is calculated: for each time v_i has debunked a fake news his credibility increases by 1, his credibility increases by 1 for each time v_i has made a neighbor share a fake news according to the bias blind spot mechanism, for each time v_i has received a debunk by a neighbor his credibility decreases by 3, if v_i has not shared a fake news his credibility increases by 1. If this new credibility is less than the older one, the strategy of v_i changes according to changes calculated in FakeNewsPropagationAlgorithm **??** setting $\phi_i \leftarrow \phi_i'$.

Algorithm 1: FakeNewsPropagation

```
begin
    foreach v_i ∈ V do
        factChecking ← random ∈ [0, 1]
        if factChecking ≤ φ_i or v_i shared a fake news then
            v_i shares the fake news
            φ_i increases by phi_i α foreach v_j ∈ δ(v_i) do
                debunk ← random ∈ [0, 1]
                if debunk ≤ φ_j then
                    // v_j debunks the news shared by v_i
                    φ'_i increases by φ_i w_{i,j} γ / ε
                    if relationshipFeedback then
                        weakens relationship {v_i, v_j} by factor β
                    end
                else
                    // The neighbor v_j shares the fake news
                    φ'_i decreases by φ_i w_{i,j} α
                    if relationshipFeedback then
                        reinforce relationship {v_i, v_j} by factor β
                    end
                end
            end
        else
            if selfDebunk then
                φ'_i increases by φ_i γ
                if relationshipFeedback then
                    Reinforce relationship with fact checkers neighbors by a
                      factor β
                end
            end
        end
    end
end
```

3.2 Experiment and results

The model has been implemented in Netlogo 6.0. In this subsection some results of simulations, executed with different settings, will be showed discussed. Is assumed that every simulation starts with 100 non-fact checkers and 100 fact-checkers.
Results of the experiment are coherent with the previously given definition of attitude polarization resulting from group interactions. As summarized in figure 1, attitude polarization seems to be faster when the starting average degree of players is higher, thanks to the group polarization which seems to occur even when the starting relations and groups are composed by heterogeneous players. When $\alpha = \gamma$ the number of players with both attitudes remains almost the same



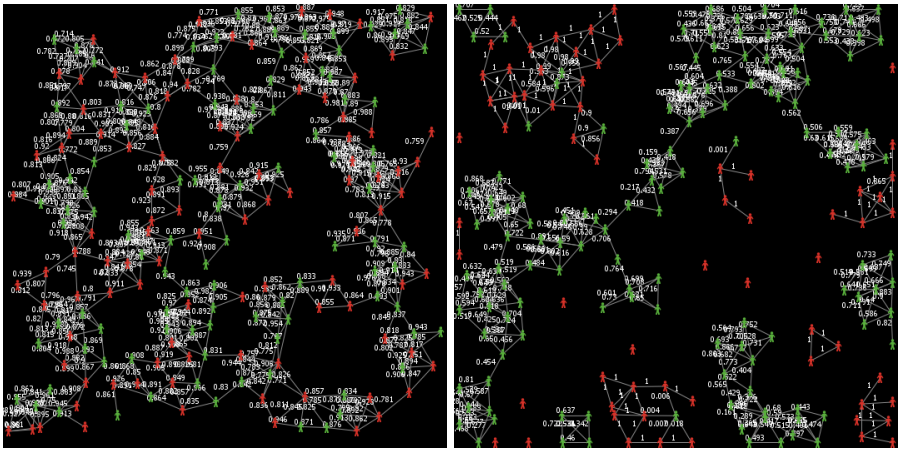
(a) $\bar{D} = 3$ (b) $\bar{D} = 5$ (c) $\bar{D} = 10$

Fig. 1: Results after 50 turns. Both the experiments started with a population equally distributed between fact-checkers and non fact-checkers. The higher the average degree, the sharper and faster the attitude polarization. Parameters are: $\alpha = 0.3$, $\beta = 0.1$, $\gamma = 0.3$, $\bar{D} = 5$, link removal enabled, relationship feedback enabled, self debunk disabled and diffidence in debunker disabled

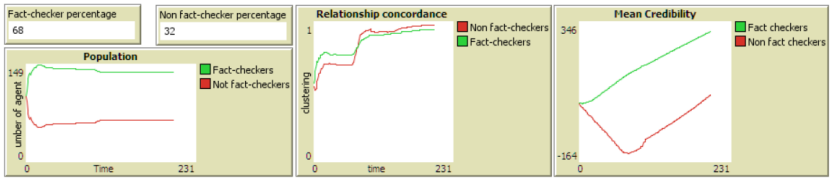
during all the simulation when $\bar{D} = 3$, when $\bar{D} < 3$ non fact-checkers number gets higher, and when $\bar{D} > 3$ there's a predominance of fact-checkers population, except when self-debunk is enabled. In this case the population is equally distributed, roughly 50% of fact checkers and 50% of non fact-checkers, only when $\bar{D} = 1$, with other values of $\bar{D}$ there's always a predominance of fact-checkers.
Another interesting behavior is the creation of homogeneous connected subgroups, also called echo chambers, highlighted by a rapid increment in relationship concordance, which is the probability that a player is linked to another player with a similar attitude, due to the removal of some links between players with different attitude whose relationship confidence got very close to zero. However relationship concordance of fact-checkers seems to be slightly lower than non fact-checkers' one. This means that very few fact-checkers are in a non fact-checkers groups. After this break up, the credibility of every isolated sub-group tends to increase.
An example of simulation is given in figure 2. In simulation view fact-checkers are in green, while non fact-checkers are in red. At the beginning $\bar{D} = 5$. A break up occurs around turn 80. In (b) the formation of echo chambers is clearly seen. There's a large group of fact-checkers in which confidence between components assumes values in (0, 1) and non fact-checkers are not present, while non fact-checkers have the maximum confidence with other non fact-checkers, but they

are connected with lower confidence to few fact-checkers, who are still present in some of these groups. If the self-debunk option is enabled confidence between fact-checkers reach the maximum value as well. In statistics (c) could be clearly seen the increase of credibility for both types of players once the break up occurs. Thus, in echo chambers players' attitude doesn't change radically, instead it will be almost the same, reinforcing previous opinions because there's no need for the player to change the strategy since the new credibility is equal or higher than the previous one.

Summarizing, if the starting average degree is high, the number of fact-checkers



(a) Beginning (b) At turn 200



(c) Statistics

Fig. 2: An example of simulation with parameters set to $\alpha = 0.3$, $\beta = 0.1$, $\gamma = 0.3$, $\bar{D} = 5$, link removal enabled, relationship feedback enabled, self debunk disabled and diffidence in debunker disabled

gets higher in the simulation, meanwhile the attitude polarization is faster and sharper for both kinds of attitudes. In subgroups composed by a predominance of non fact-checker players there could be few fact-checkers, but in groups with a predominance of fact-checkers, players with an opposite attitude are not present. Credibility of non fact-checkers gets higher exclusively once they break up the relationship with a large group of fact-checkers, this result is coherent with the notion of "naive realism", but if the relationship feedback is disabled, their aver-

age credibility becomes less predictable, as it seems to suddenly increase in some points as showed in Figure 3. These are the most interesting cases and results found during experimentation. Other experiments could be done by changing parameters and enabling switches already implemented in Netlogo.

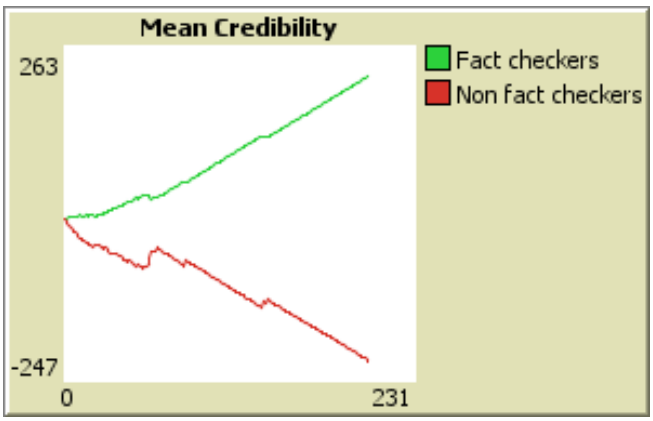


Fig. 3: This is the situation after 200 turns, with parameters set to $\alpha = 0.3$, $\beta = 0.1$, $\gamma = 0.3$, $\bar{D} = 5$, link removal enabled, relationship feedback disabled, self debunk disabled and diffidence in debunker disabled. The mean credibility of non fact-checkers becomes less predictable, as it increases sometimes, slightly decreasing fact-checkers' credibility at the same time.

4 Conclusion

The problem of fake-news and misinformation diffusion has increased since the birth of the internet and it has become crucial after the advent of social media, enough to be inserted in today's global risks by WEF (World Economic Forum) already in 2013 [15]. After the exploration of cognitive biases involved in evaluation of a news it has been introduced game theory and a model based on it able to simulate, more or less, behavioral polarization and the emerging of echo chambers in a complex social network if cognitive bias and debunk effectiveness are equal. It is clear that the multidisciplinary nature of the issue makes impossible to understand the entire question exclusively through game theory. Cognitive biases involved in news judgement may be more than the ones described in this report and discovered today. Simulation's results, however, seemed to confirm some of the psychological and sociological studies about naive realism and its consequences in social networks, but they need to be compared with a real situation in order to be better understood. Unfortunately fake news and naive realism seem to be well radicalized in people and is important to understand that this kind of behaviors existed even before the emerging of new technologies, which

amplified this phenomenon. To reduce this amplification is important to educate users, informing them about fake-news mechanisms, helping them to understand the differences between an authoritative source or non authoritative source, enhancing the effectiveness of debunk and reducing the backfire effects. As stated by Cook and Lewandowsky [12] is important to increase people's familiarity with the facts, providing few but clear arguments, because in the absence of better explanations, people opt for the wrong one, this means that misconceptions must be explained in the first phases of a debunk, in order to reduce backfire effects. The work presented in this report could be extended in future, adding new features for more realistic interactions between people, trying other settings, introducing new behavioral attributes for players or modifying the news spreading algorithm with other biases existing in sociological and psychological studies, but ignored in this model or not discovered yet.

References

1. Robert Axelrod. *An Evolutionary Approach to Norms.* American Political Science Association, 1986.
2. Hunt Allcott and Matthew Gentzkow. *Social Media and Fake News in the 2016 election.* Journal of Economic Perspectives, 2017
3. Ross Lee and Ward Andrew *Naive realism in everyday life: Implications for social conflict and misunderstanding* 1996 In T. Brown, E. S. Reed and E. Turiel (Eds.), Values and Knowledge (pp. 103135). Hillsdale, NJ: Erlbaum.
4. Emily Pronin, Daniel Y. Lin, Lee Ross *The Bias Blind Spot: Perceptions of Bias in Self Versus Others* 2002: Personality and Social Psychology Bulletin. 28 (3): 369381.
5. Robinson, Robert J.; Keltner, Dacher; Ward, Andrew; Ross, Lee *"Actual versus assumed differences in construal: "Naive realism" in intergroup perception and conflict"* 1995: Journal of Personality and Social Psychology. 68 (3): 404417.
6. Lee Ross, Constance A. Stillinger *"Psychological barriers to conflict resolution"* 1988: Stanford Center on Conflict and Negotiation, Stanford University,
7. Katherine Hansen, Margaret Gerbasi, Alexander Todorov *People Claim Objectivity After Knowingly Using Biased Strategies* 2014
8. Charles G. Lord, Lee Ross, and Mark R. Lepper *Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence* 1979: Journal of Personality and Social Psychology vol. 37, No. 11, 2098 - 2109
9. Mackie, Diane M.; Worth, Leila T.; Asuncion, Arlene G. *Processing of persuasive in-group messages.* 1980: Journal of Personality and Social Psychology, Vol 58(5), May 1990, 812-822.
10. D. K. Freedheim (Ed.) *History of psychology. Vol. 1 of I. Weiner (Ed.) Comprehensive Handbook of Psychology. New York: Wiley* .
11. Engin Bozdag *Bias in algorithmic filtering and personalization* 2013: Ethics and Information Technology. 15 (3): 209227.
12. Cook, J.; Lewandowsky, S. *The Debunking Handbook* 2011: St. Lucia, Australia: University of Queensland
13. Gary William Flake *The Computational Beauty of Nature*
14. David Easley and Jon Kleinberg.*Networks, Crowds, and Markets: Reasoning about a Highly Connected World.* 2010: Cambridge University Press
15. World Economic Forum Globl Risks 2013 *http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/*