Lecture with Computer Exercises:
Modelling and Simulating Social Systems with MATLAB

Project Report

Investigating the Influence of Fake News on Opinion Formation

Yannick Bertschy, Mario Blatter, Guido Gandus, Aiping Yao

Zurich
December, 2017

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Influence of Fake News on Opinion Formation

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Bertschy	Yannick
Blatter	Mario
Yao	Aiping
Gandus	Guido

With my signature I confirm that
- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date

Zurich, 13.12.17

Signature(s)

[illegible]
Aiping Yao
[illegible]
Guido Gandus

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Agreement for free-download

We hereby agree to make our source code of this project freely available for download from the web pages of the SOMS chair. Furthermore, we assure that all source code is written by ourselves and is not violating any copyright restrictions.

Yannick Bertschy	Mario Blatter	Guido Gandus	Aiping Yao

Table of Contents

1. Abstract

This paper investigates the influence of fake news on a society and specifically on opinion formation. This work builds on the model presented by P. Holme et. al. in [1] on nonequilibrium phase transition in the coevolution of networks and opinions, with one substantial extensions. To be more specific we apply an additional, external influence, we refer to as fake news.
We study the dynamics of opinion formation within the network both in the absence and presence of fake news. Two strategies are proposed to investigate the most efficient business model to influence the opinion formation in a society with two opinions using a fake news input.
The first strategy specifically targets the most famous person in the society, i.e. the vertex with most connections in the network. The second strategy is to affect more but less "popular" people, i.e. vertices with a lower degree but a cumulative degree equal to the one of the most connected vertex.
Our results show that based on the model used in this work, the second strategy is more efficient than the first one.

2. Individual contributions

The whole project was done in a cooperative manner among all the group members.
To be more specific, Mario Blatter mainly worked on the implementation and evaluation of the original model, Aiping Yao implemented - with the help of Guido Gandus - the extended one. Yannick Bertschy, partly along with Guido Gandus, wrote the better part of this report.

3. Introduction and Motivation

In view of recent discussions about the importance of so-called fake news stories having made appearance on social networks such as Facebook or Twitter, questions about their impact have been brought up. The undoubtedly most popular example are the 2016 US elections where this problem was controversially discussed in both media and politics.

Fake news can refer to various different things. In this work we understand it as pieces of information that are deliberately spread via - mainly online - media outlets in order to influence a people's opinions. This aligns with the definition used in [2].
To better understand the influence of fake news on a society and specifically on opinion formation within it, we implement a model based on the one presented in the paper by P. Holme et. al and extend it. They described a model that combines opinion dynamics with assortative network formation [1]. It means that opinions might change as a result of the network in place and at the same time the network evolves depending on the opinions in it. It is referred to as the coevolution of network and opinion. This is likely to happen in the real world and often leads to homophily in a society [3].
Boosted by the increasing importance of information and opinion spreading via online media in the US elections this is an ongoing topic in research [4,5].

The main goal of our project is to reconstruct this agent-based model in order to investigate the dynamics of the original model when varying the impact of network formation and opinion formation.
In addition, we extend the model such that we see the influence of external fake news implemented by us. Further, we compare different strategies on implementing fake news and with the help of a cost function we find out which is the most efficient one.

In the following we describe the model by P. Holme et al. before we add the extended version.
After that, we analyse and discuss the behaviour of the extended model and try to investigate how fast and efficient we can drive opinions towards a desired one.

4. Description of the Model

4.1 Original model

As mentioned in the previous chapter, the model presented in this project is based largely on the one described in the paper by P. Holme and M.E.J. Newman [1], but with the additional implementation of an external influence.
The model figures N individuals as vertices. Each of these vertices (i) can have an arbitrary number of pairwise connections to another vertex representing friendship or acquaintance. The total number of connections in the network is denoted by M. Hence, each individual has 2*M/N edges on average.

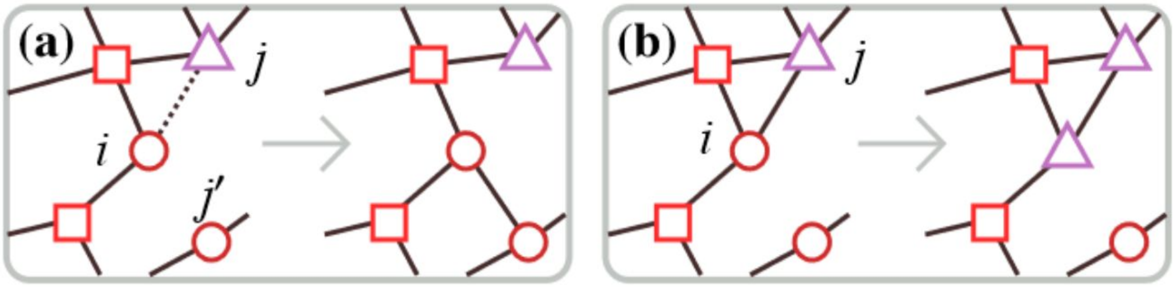

Figure 1. Representation of the two possible network modifications that can take place in the base model [1].

Each individual holds one out of G possible opinions on a topic (g(i)). For our addressed topic G = 2 and corresponds to the two main candidates from the 2016 US elections, thus G = {Hillary Clinton, Donald Trump}. As the starting conditions for the model, we distribute the total number of connections M at random and assign each vertex an opinion in G. Note that, for the sake of simplicity, we allow for multiple and self connections as was done in the original model.
The dynamics of the system follow two rules:

1. Pick a vertex (i) at random. With the probability φ a random connection to a vertex of a different opinion is changed to a random vertex of its same opinion g(i) (see figure 1 a)).
2. With probability 1-φ the vertex (i) adopts the (different) opinion of one of its neighbouring vertices (see figure 1 b)).

4.2 Extended model

This report focuses on the effect of fake news on opinion formation. We want to investigate which of two proposed strategies is the more efficient to significantly shift the initial (random) distribution of opinions in a society.
The implementation of fake news is superimposed onto the dynamics described in the previous section. At each iteration, the network can undergo a change following these three possible events:

1. A random edge is moved to lie between two individuals whose opinions align (equivalent to point 1 in 4.1).
2. The opinion of an individual is changed to the one of a random neighbour (equivalent to point 2 in 4.1).
3. One or more vertices change their opinion with a probability of β due to the fake news, a pre-defined opinion of our choosing.

Note that events 1 and 2 are mutually exclusive as before, i.e. cannot occur at the same time as the first occurs with probability φ and the second with 1-φ. And that event 3 is subject to the probability variable β and can occur independently from the others. We implement the

latter to investigate the most efficient business model - or strategies - to target a society with fake news.
Two different strategies are investigated. The first one - depicted in figure 2 - where the most famous person in the society is influenced by our fake news. This is defined as the vertex with the highest degree of the opposing camp. Since it has the most connections, one could think that the effect on opinion formation would occur at the highest rate.

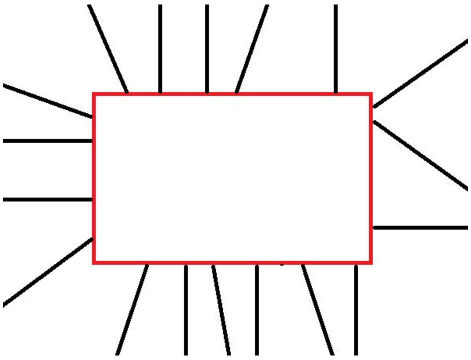

Figure 2. Illustration of strategy 1.

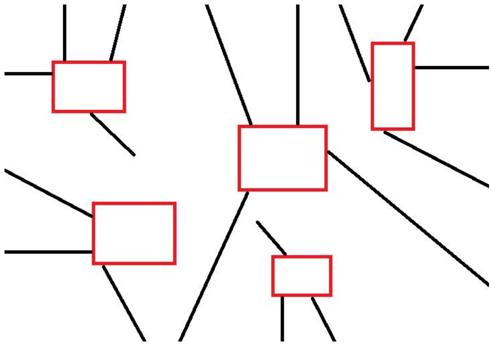

Figure 3. Illustration of strategy 2.

In the second scenario, shown in figure 3, we influence more people which are less ”popular”, i.e. each individual has fewer connections than the most famous person. This is equivalent to saying that we choose vertices which have a lower degree. In this scenario the individuals (including their number) to be affected are chosen randomly among the vertices with different opinion. The only condition being that their cumulative degree is equal to the one of the most connected vertex at that iteration.
Both scenarios will have the same amount of connections at the beginning and we simulate the two strategies in Matlab in order to seek the rate of spread of the fake news. In this view, the cost for the business model is determined by the number of connections we influence during each iteration. The simple approach we take is that cost scales linearly with the number of connections. Given an initial budget of connections, at each time step we invest according to the strategy we selected at the beginning.
We simulated both scenarios under varying population size N, number of connections M, initial budget of connections and probability of being affected β. If not specified, we assume φ=0.9. Contrary to Holme we fix the number of iterations at 1000, as the time frame to influence people before an election is limited. The best strategy is subsequently assessed by comparing the number of iterations required to get 80% of the population to agree on the same opinion.

5. Implementation

5.1 Original model

The realization of the model consists of a model initialization followed by iterations for a number of timesteps specified below. This is executed when running the script “*Without_fake_news.m*”.

To initialize the framework an array representing the opinions held by each individual was created. Opinions were assigned at random. Then, a connectivity matrix was set up to represent the initial state of the network. This is implemented in the function file "*initialize.m*". Along the main model described in "*opinion_change_model.m*" these two instances were modified with each individual iteration according to the two criteria specified in section 4.1. We use a random value in [0,1] and compare it against our defined probability φ. In a first attempt we ran 10^6 iterations in order to reach consensus state. In an improved version an abort criterion was introduced. It would stop the iteration of the model when there had not been any change in the population sizes for the last 10^4 timesteps.

5.2 Extended model

We define the population size N, the number of connections M and the opinion groups G. The M edges are randomly distributed in the network and stored in a connectivity matrix. Opinions are assigned to each of the N individuals. Three different computer simulations are performed for the same initial configuration of the network. The first one simulates the dynamic of the society in the absence of fake news, the second and the third one implement the presence of fake news for the two business models described in the previous section. As a starting point, both models have the same budget of total connections to spend.

Then, on each step we pick one vertex *i* at random. If the degree of connections is zero we proceed to the next step, else we apply the same two rules used by P. Holme et al. explained in section 4.1. In parallel, we search for the individual with the highest number of connections from the set of individuals having a different opinion than the targeted vertex. We store its number of connections as the current budget to spend. In the first scenario we invest the sum in the individual with the highest number of connections, in the second into many individuals with an equivalent number of connections.
To be more specific, in each iteration we "add" fake news by changing the target's opinion with a probability of β. All fake news related parameters (β, strategy, budget and whether to add it or not) are defined as a structure variable 'Fake'. The person chosen to be affected depends on the strategy as described above.
Then, the percentage of the population for each opinion group is calculated as a measure of the effect of the fake news.
The whole procedure is performed by running "*With_Fake_News_Final.m*" . The extended model is defined as a function by *"extended_model_v2.m*", with the implementation of the fake news being performed through the function of "*fake_news_effect.m*".

6. Simulation Results and Discussion

6.1 Original model

In order to compare the results of our implementation and the ones obtained by Holme, we decided to replicate the probability of the population sizes for three different transition probabilities φ (0.04, 0.458, 0.96). We use the same number of connections M (6400) and number of individuals (3200) as [1].
Our data were binned logarithmically as was stated in the paper [1]. Note however that the binning appears to be rather arbitrary (see figure 4). It was aimed at replicating the binning by reading out the corresponding values on the x axis as exact as possible as this influences how the data look in the end.

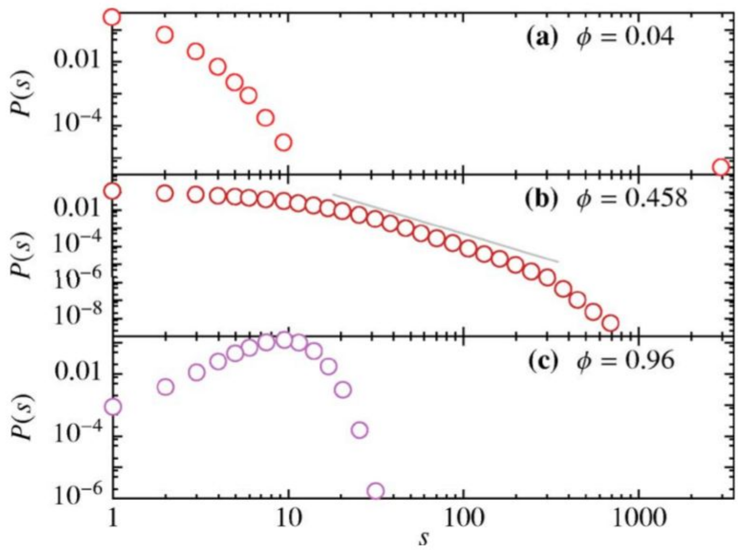

Figure 4. Histogram of community sizes and their respective probability of occurrence according to Holme [1].

6.1.1 Fixed number of iterations

Figure 5 shows the results of our replication of Holme's model.
Note the occurrence of intermediate cluster sizes in the range of 20 to 3000 people for φ = 0.04 and the less broad distribution of cluster sizes for φ of 0.458. This behavior indicates that the replication of the model was not successful in this first attempt.
The assumption that a duration of 10^6 iterations was enough to reach consensus state for at least some runs turned out to be wrong. This then lead to the occurrence of cluster sizes that would vanish with further iterations.

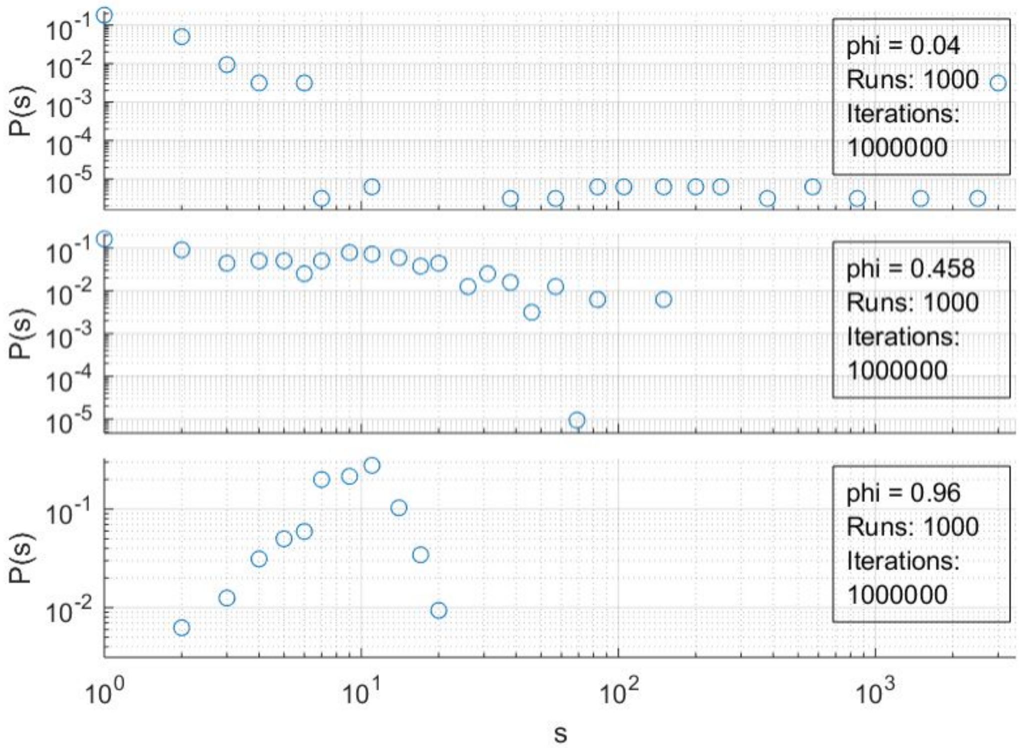

Figure 5. Probability density function of the possible cluster sizes with P(s) denoting the probability of a cluster's occurence in the consensus state. Replication of Holme's model using N = 3200, M = 6400 and G = 320.

6.1.2 Varying number of iterations using abort criterion

Due to the problem mentioned above the slightly different approach of using an abort criterion was applied. This means that for each run, iteration would stop when the cluster sizes in the network had not changed for the last 10^4 timesteps.
Figure 6 shows the outcome of our model iterations when using said abort criterion.
Note: Due to limited computing power and time this part was performed using the average over a limited number of runs specified in the figures.

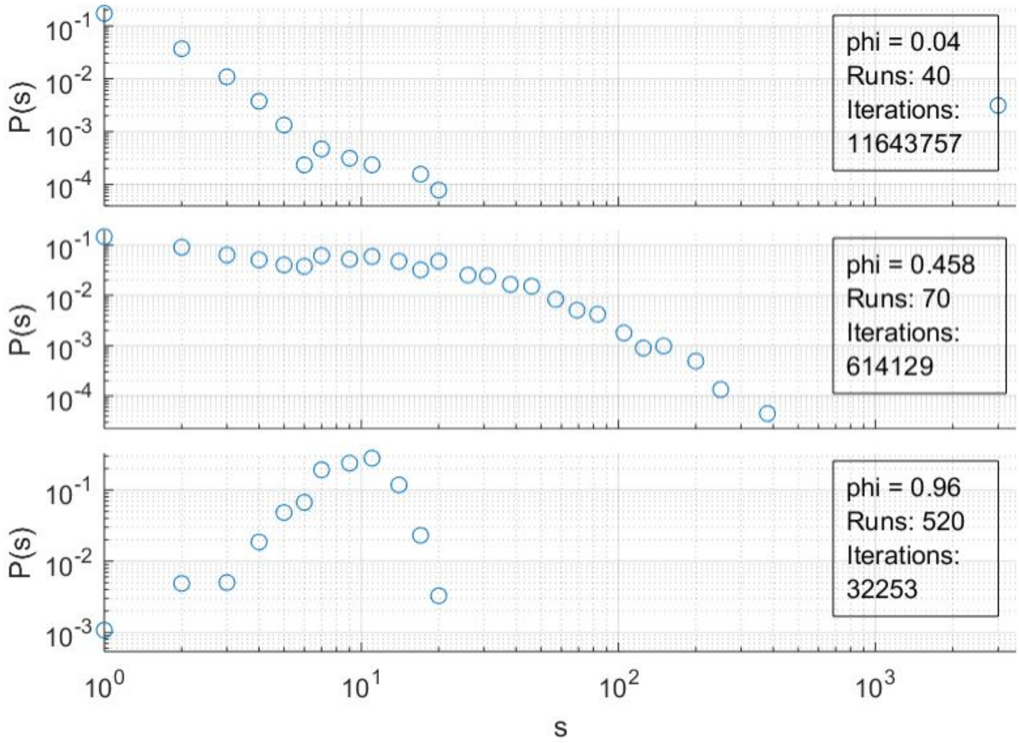

Figure 6. Probability density function of the possible cluster sizes with P(s) denoting the probability of a cluster's occurence in the consensus state. Replication of Holme's model using N = 3200, M = 6400 and G = 320. The iterations were terminated as soon as the abort criterion was fulfilled. The iteration count refers to the average number of iterations undertaken until the abort criterion was met.

Clearly, there exist two types of final networks, i.e. cluster distributions. For small and large φ, most population sizes are small with the exception of one giant cluster that may form for φ = 0.04. For the intermediate probability also larger cluster sizes may develop. We effectively reproduce the "phase transition behavior" described by Holme.
At the lower end of the cluster sizes the probability is comparable to the the original work. The most notable deviations of this replication are the orders of magnitude towards the upper end of the x axis. The giant cluster forming under φ = 0.04 is more than 100 times more likely to occur than in Holme's results. It has to be noted that the number of runs is substantially lower and thus the statistical sample is not very large. This means that the coincidental occurrence of a big cluster has a higher impact than when we average over more runs. It is argued here that this is the reason for the different occurrence of this cluster. This could be further investigated with more computing power and time.

The above results show that the reproduction of the model was successful to a large degree and can serve as a useful starting point for variations.

6.2 Extended model

The cluster sizes of the two opinion groups (expressed as a percentage of the total population) are studied and compared with different parameters. The effect of the two strategies are compared specifically for each of the resulting situations.
In the consideration of the running time of the simulation, for every study we keep the iteration time to be 1,000 times. The related parameters which we can manipulate and change are defined in Table 1.

Table 1: Parameters for the model.

Strategy 1	Affect one famous person
Strategy 2	Affect several people
G1	Opinion of group 1
G2	Opinion of group 2
Budget	Total connections allowed to affect
Beta	Probability of a person been affected by the fake news
Phi	Probability of changing to the same opinion
N	Population Number
M	Connection Number

The first scenario simulates a society with population size of N=3,200 with M=10,000 connections between the people. These people have a large probability of being influenced by fake news (β = 0.8). The available budget is 5,000. And the goal of the use fake news is to make group 1 greater.
Figure 7 shows the results for this scenario. As expected, it is clear that the fake news have a huge effect on the population percentage. It also becomes apparent that the goal of reaching the 80% line will never be fulfilled with neither strategy 1 nor strategy 2. This is because after 510 iterations, the cost has reached the budget with strategy 1, thus no fake news can be added anymore, and the influence stops after 61%. With strategy 2, one can see the advantage of it: Even though the goal is also never achieved, one can influence about 74%.

In the second scenario (see figure 8), we double the budget to 10,000. As expected, the effect of fake news will increase as well and more people can be affected. With strategy 1, the goal of reaching the 80% border will probably be reached after 1,000 iterations.
While with strategy 2, the target of 80% population percentage of G1 is achieved after only 550 iterations whereby one has spent the entire budget after 770 iterations. After that, no fake news will be added to the society.

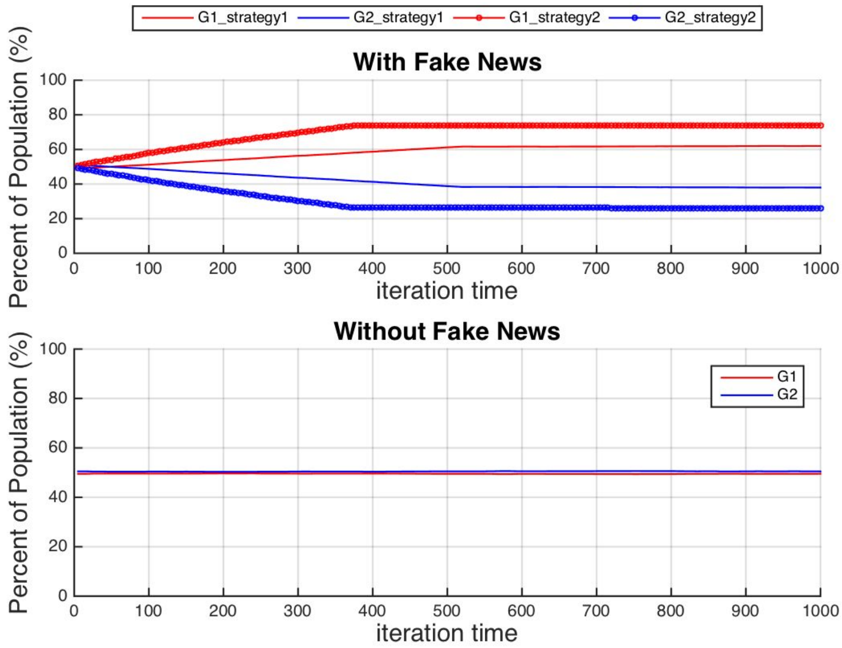

Figure 7. Top: Comparison of fake news effect on the percentage of opinion group populations with initial condition of N=3,200, M=10,000, β=0.8, Budget=5,000, φ=0.9. Bottom: Percentage of opinion group populations without the fake news under same initial conditions as on top.

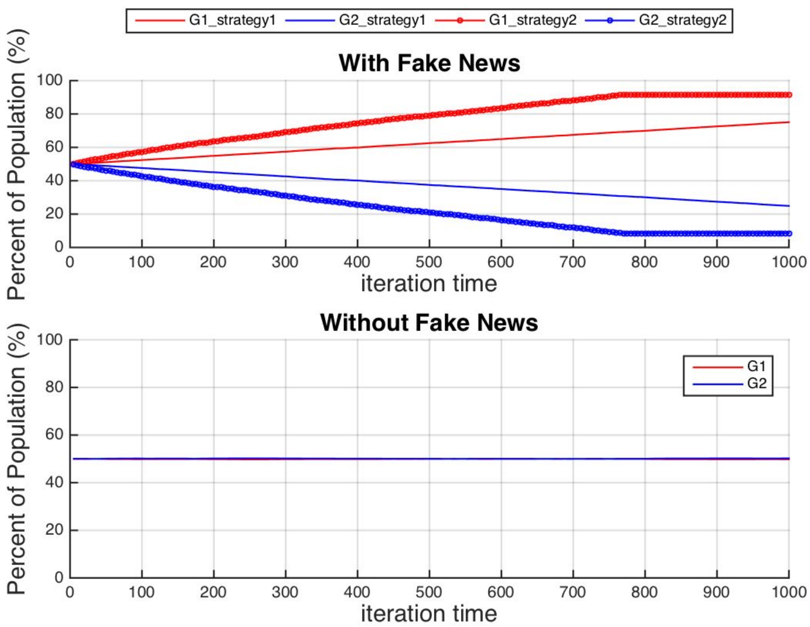

Figure 8. Top: Comparison of fake news effect on the percentage of opinion group populations with initial condition of N=3,200, M=10000, β=0.8, Budget=10,000, φ=0.9. Bottom: Percentage of opinion group populations without the fake news under same initial conditions.

Then, we construct a society with the same population but half the connections M among people. Under the same budget, β and φ. As shown in figure 9, with strategy 1, although the society is less related to each other than before, the rate of change is the same as for the well-connected group. This might be surprising, since one would expect a slower spreading of fake news in such a group, since the average degree is lower.
However, there are differences concerning strategy 2. Executing the implementation with strategy 2 gives very satisfying results. It is successful in suppressing the other opinion such that after approximately 940 iterations the entire society has exactly the same opinion. The 80% population is reached for our target group G1 after 460 iterations which is also faster than the society with more connections among people. This result is not straightforward but reasonable, as in each iteration, the number of people affected in this case will be higher than in the case with more connections.

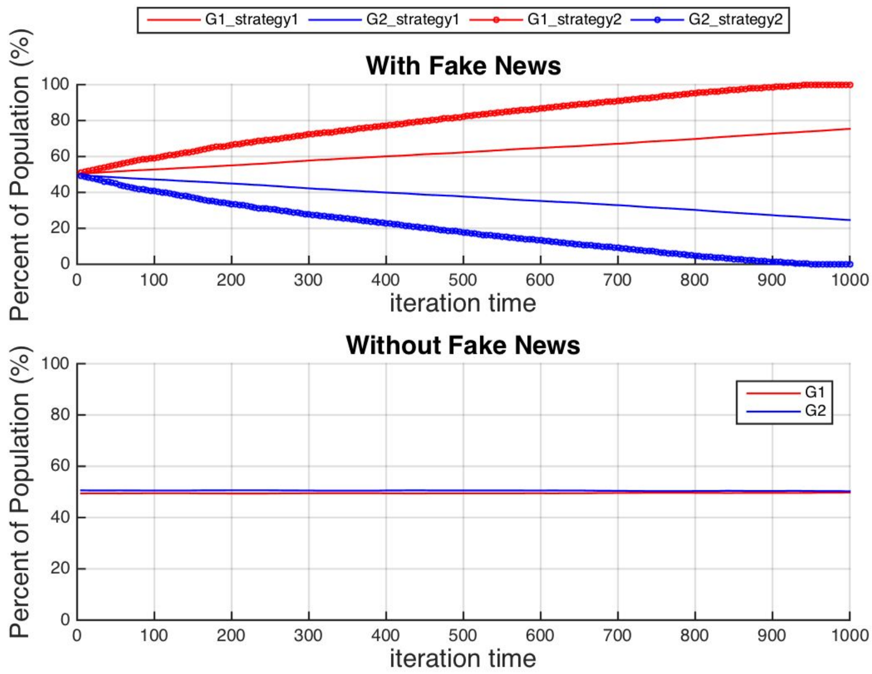

Figure 9. Top: Comparison of fake news effect on the percentage of opinion group populations with initial condition of N=3,200, M=5,000, β=0.8, Budget=5,000, φ=0.9. Bottom: Percentage of opinion group populations without the fake news under same initial conditions.

The scenario shown in figure 10 deals with a smaller group within a population of N=1,600 but again with the amount of connections M=10,000 and a budget of 10,000.
As one would expect, with both strategy 1 and strategy 2, it will be faster to reach the 80% aim than with the large population society. But strategy 2 is still better than strategy 1, with the difference in the number of iterations needed to reach 80% being smaller than before. And the final population distribution for both strategies is better than in a larger population society: 90% against 64% for strategy 1 and 92% against 74% for strategy 2.

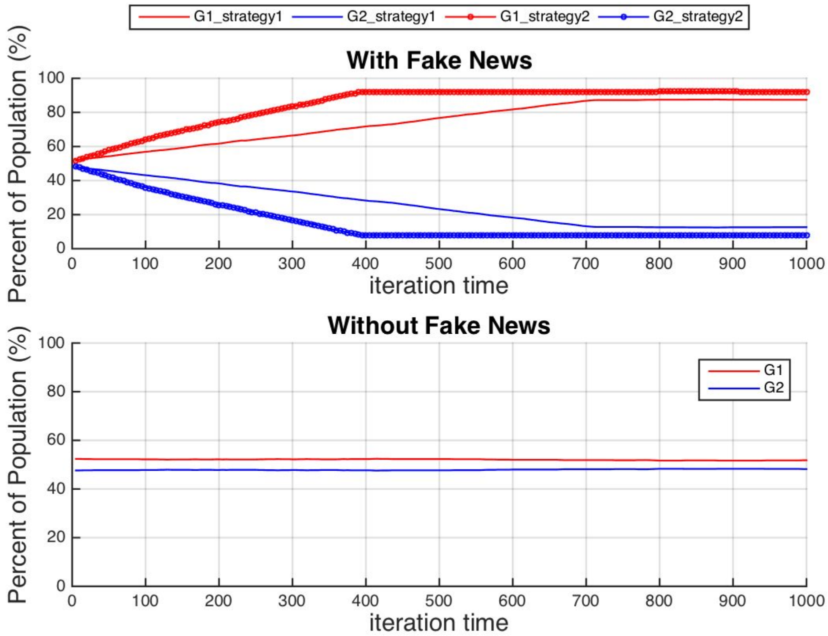

Figure 10. Top: Comparison of fake news effect on the percentage of opinion group populations with initial condition of N=1,600, M=5,000, β=0.8, Budget=5,000, φ=0.9. Bottom: Percentage of opinion group populations without the fake news under same initial conditions.

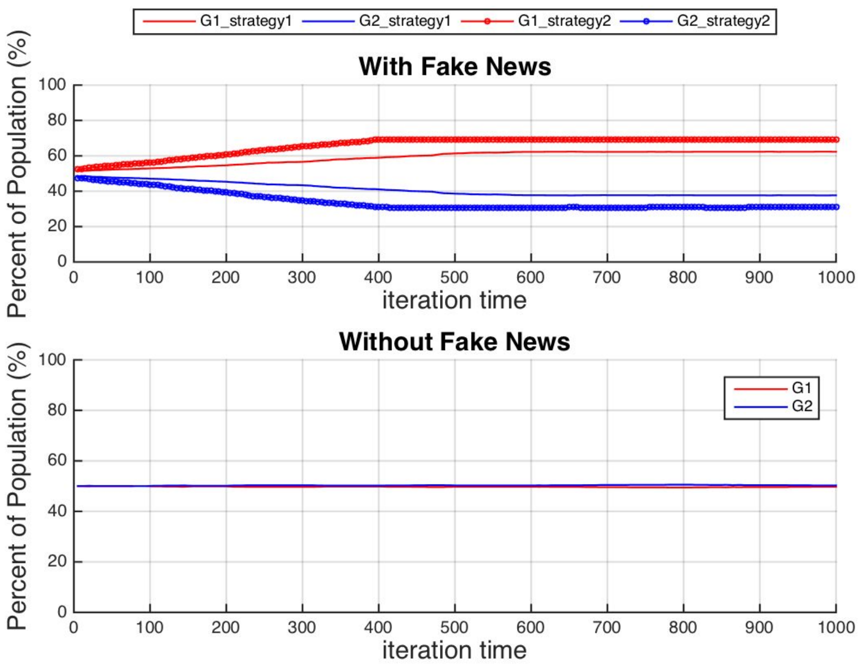

Figure 11. Top: Comparison of fake news effect on the percentage of opinion group populations with initial condition of N=1,600, M=5,000, β=0.3, Budget=5,000, φ=0.9. Bottom: Percentage of opinion group populations without the fake news under same initial conditions.

In the last scenario which we investigate, we lower the probability of being affected by fake news to β=0.3. This is chosen for a group number of 1,600. As expected, it is more difficult to change the opinion in the society. However, it is still possible to reach a level of almost 70% with strategy 2. With strategy 1, however, one can only get a proportion of 62% on target group G1.

From the five scenarios discussed above, we can conclude that - using our model - it is always more efficient to target fake news on several people instead of just on the most famous person.

So far, we assumed φ=0.9. For the next experiment we would like to vary this parameter in order to find out if there are any differences. Comparing the plots in figure 12 and figure 11, one can conclude that φ has no or negligible influence on either of our strategies.

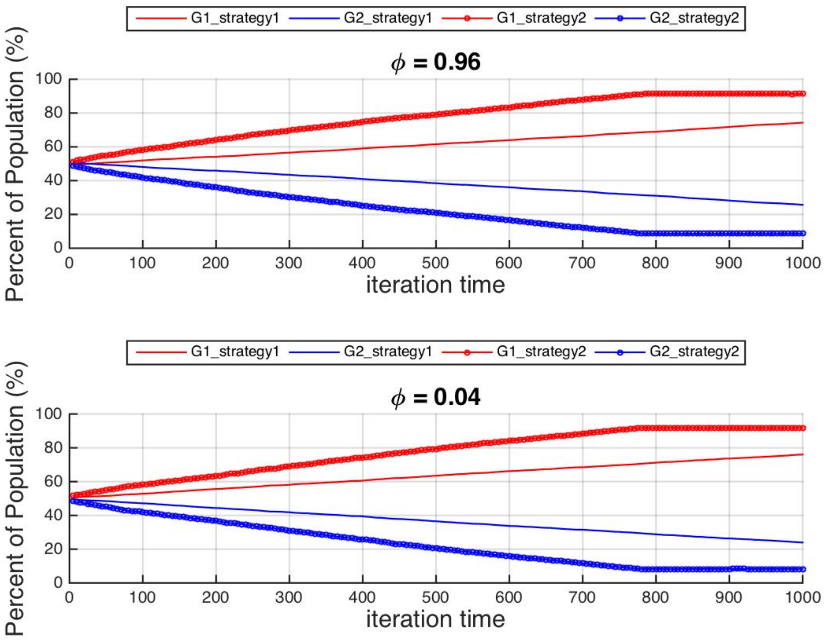

Figure 12. Top: Comparison of fake news effect on the percentage of opinion group populations with initial condition of N=1600, M=5000, β=0.3, Budget=5000, φ=0.96. Bottom: Comparison of fake news effect on the percentage of opinion group populations with initial condition of N=1600, M=5000, β=0.3, Budget=5000, φ=0.04.

To have a more detailed study of how budget will change the fake news effect, we choose a society with a population of N = 3200, M = 10000, and β=0.8, while changing the budget from 1000 to 15,000. As shown in figure 13, it does not matter how much money one would like to spend in order to influence a society, the strategy 2 is always better than strategy 1. For suppressing one opinion completely with strategy 2, the budget needed to be over 12000. But for strategy 1, the maximum percentage will be around 75% with the iteration of 1000 times. Therefore, the maximum budget one should spend on strategy 2 is around

8000. Anything above strategy 1 cannot compete in this period of time. Otherwise one would need to run more iterations to reach the 80% goal, which means more time in the reality.

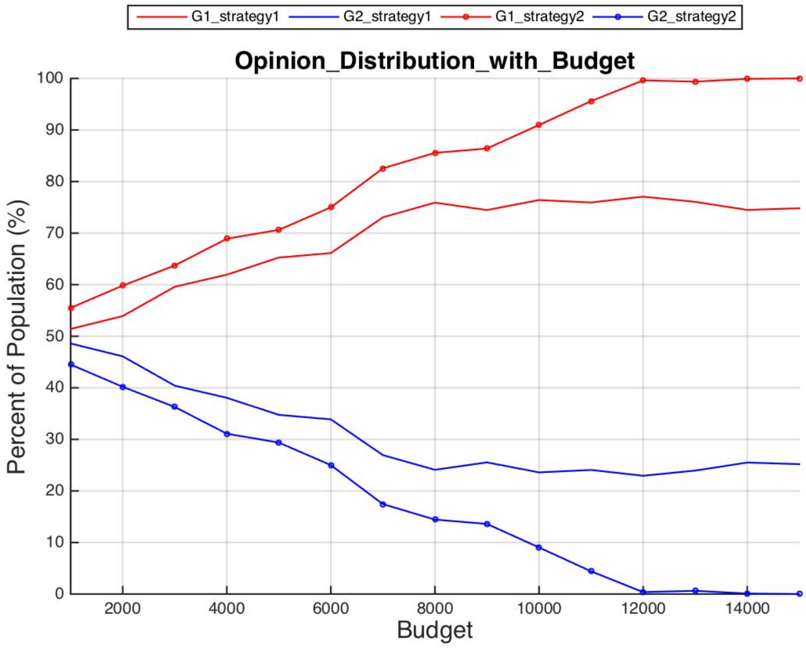

Figure 13. Comparison of fake news effect on the percentage of opinion group populations with N = 3200, M = 10000, β=0.8, but different budgets changing from 1,000 to 15,000.

In the next study, we keep other parameters constant (N = 3,200, budget = 10,000, and β=0.8) and change the number of connections M from 5,000 to $5x10^5$. Figure 14 shows that changing the number of connections within a group also affects the behaviour of our model. For a larger number of connections one cannot influence the opinion in the society. For a number of connections above 10^5 it does not make sense to invest in fake news with the given budget, since it will not have a significant influence on the society. Alternatively we would need to invest more money.

In the end we will look more closely at the effect of different population sizes. As shown in figure 15, with relatively small population sizes, strategy 1 and strategy 2 tend to have a comparable effect. However, when the population size increases, the strategy 2 tends to have a better performance, and the difference between strategy 1 and strategy 2 also increases with population size. Which means the larger population size is for a society, the strategy of affecting a famous person is less efficient compared to targeting several “ordinary” people.

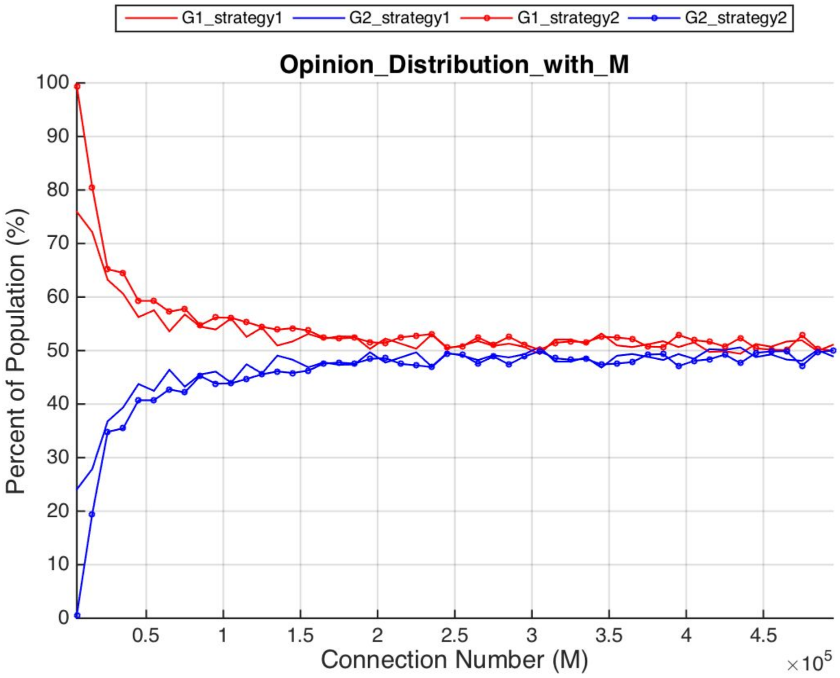

Figure 14. Comparison of fake news effect on the percentage of opinion group populations with N= 3,200, budget = 10,000, β=0.8. Number of connections changing from 1,000 to 500,000.

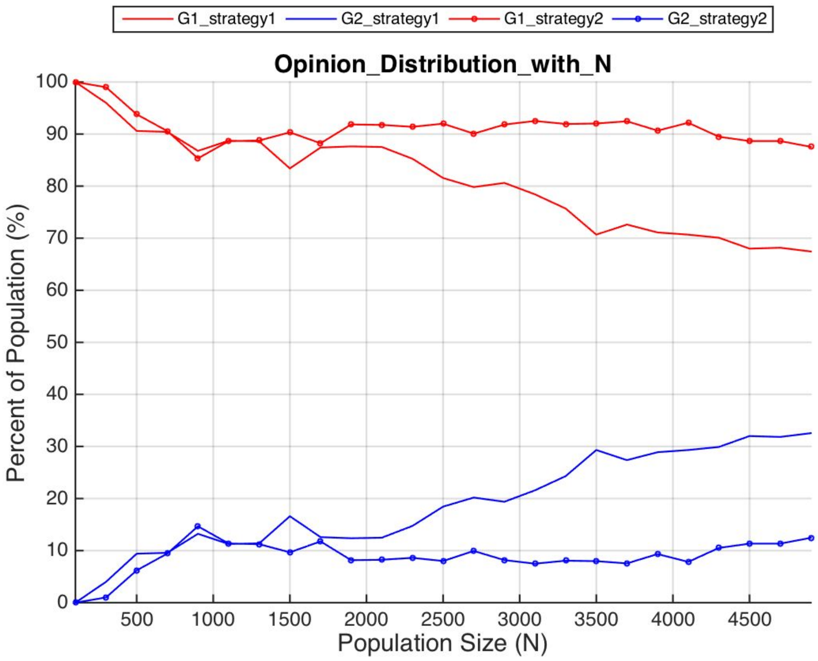

Figure 15. Comparison of fake news effect on the percentage of opinion group populations with budget = 10,000, M = 10000, β=0.8. Population size changing from 0 to 5,000.

7. Summary and Outlook

In the first part of this paper, we presented the model proposed by Holme, P. et al in [1]. After having successfully replicated the results obtained in [1], we proceeded to the simulation of the model extended with an external influence we call fake news.
Two strategies are proposed to implement the fake news effect. The first strategy specifically targets the most famous person, whereas in the second strategy we affect more but less “popular” people.
Both scenarios are simulated under variation of the population size N, the number of connections M, the initial budget of connections, the probability φ and the probability of being affected β.
The results of the simulations lead to the following conclusions. With relatively small population sizes, strategy 1 and strategy 2 tend to have similar effect. When the population size increases, the strategy 2 tends to have a better performance, and the difference between strategy 1 and strategy 2 also increases with population size.

Note that our model does not take into account the temporal spreading of information. For the sake of simplicity it is assumed that the total “amount” or exposure to fake news is constant throughout time. But in reality the number of fake news stories fluctuates and will peak at some points. This remains to be investigated in the future.
Also, the viral and exponential spreading of stories on social media is not implemented and could also be used as an extension of our model.
Another interesting topic would be the investigation of the behaviour in a network where no coevolution exists. And with social networks taking measures to detect and neutralize fake news it will become increasingly difficult for it to have an effect on opinion formation. This would have to be taken into account when setting the probability β of being able to change a person’s opinion.

8. References

[1] Holme, P. and Newman M.E.J (2006). *Nonequilibrium phase transition in the coevolution of networks and opinions.* Physical review E74(5), 056108.
[2] Hunt, A. and Gentzkow, M. (2017). *Social Media and Fake News in the 2016 Election.* National Bureau of Economic Research, 23089.
[3] McPherson, M. et. al. (2001). *Birds of a Feather: Homophily in Social Networks.* Annual Review of Sociology Volume 27. Pages 415-444.
[4] Yosh Halberstam & Brian Knight (2016). *Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter*. Journal of Public Economics.
[5] Buccafurri, F. et. al. (2015). *A new form of assortativity in online social networks*. International Journal of Human-Computer Studies, Volume 80. Pages 56-65.

9. Appendix

The Matlab code, the corresponding data, plots and figures as well as this report can be found under the following link:
https://github.com/AipingYao/How_Fake_News_Affect