Supplementary Information for "Mitigating opinion polarization in social networks using adversarial attacks"

Michinori Ninomiya[1], Genki Ichinose[1*], Katsumi Chiyomaru[2] and Kazuhiro Takemoto[2]

[1]Department of Mathematical and Systems Engineering, Shizuoka University, Hamamatsu, 432-8561, Japan
[2]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan
* Corresponding author (ichinose.genki@shizuoka.ac.jp)

September 19, 2024

S1 Derivation of Eq. (4) used for polarization control

S1.1 Derivation for target state ± 0

In the simulation model of the echo chamber phenomenon [7], the agent's opinion dynamics are represented by the following differential equation.

$$\dot{x}_i = -x_i + K \sum_{j=1}^{N} A_{ij}(t) \tanh(\alpha x_j) \tag{S.1}$$

For this differential equation, a difference approximation using the Euler method yields:

$$x_i(t+dt) = x_i(t) + dt \left(-x_i(t) + K \sum_{j=1}^{N} A_{ij}(t) \tanh[\alpha x_j(t)] \right) \tag{S.2}$$

where dt is the appropriate time step width.

From the perspective of suppressing opinion polarization, the objective is to maintain the absolute value of each agent's opinion x in a small state while keeping its sign. When $x_i^* \geq 0$, x_i^* is targeted to $+0$. When $x_i^* \leq 0$, x_i^* is targeted to -0. Considering adversarial attacks that bring $x_i(t+dt)$ closer to the target state, polarization mitigation is applied by minimizing the energy E, defined as the correlation coefficient between the observed opinion state and the target opinion state.

$$E = \frac{1}{N} \sum_{i=1}^{N} x_i^* x_i(t+dt) \tag{S.3}$$

Here we consider minimizing E by varying the weights of the links in the network. Specifically, we employ gradient descent to introduce perturbations into the adjacency matrix at each time step. Assuming that the weights of self-loops and links that do not exist in the original network are ignored, we consider adding perturbations to the weights of links for node pairs (i, j) with $i \neq j$ and $A_{ij} \neq 0$ at time step t as follows.

$$A_{ij}^*(t) = A_{ij} - \epsilon \frac{\partial E}{\partial A_{ij}} \tag{S.4}$$

Here, ϵ is a small positive value. Focusing on a specific node pair (i, j), from Eqs. (S.2) and (S.3), the gradient $\partial E / \partial A_{ij}$ in Eq. (S.4) can be expressed as follows.

$$\frac{\partial E}{\partial A_{ij}} = \frac{x_i^*}{N} \frac{\partial \{x_i(t+dt)\}}{\partial A_{ij}} = \frac{x_i^*}{N} dt K \tanh[\alpha x_j(t)] \tag{S.5}$$

In this case, the final equation for adding perturbations to the network is as follows:

$$A_{ij}^{adv}(t) = \quad A_{ij}(t) - \epsilon \frac{x_i^*}{N} dt K \tanh[\alpha x_j(t)] \tag{S.6}$$

Also, in Eq. (S.5), since $dt > 0, \quad N > 0, \quad K > 0$ and $\quad \alpha > 0$, the optimal maximum value norm of the gradient is given by $\text{sign}[\partial E / \partial A_{ij}(t)] = \text{sign}[x_i^* x_j(t)]$. When using this, Eq. (S.6) is transformed as follows.

$$A_{ij}^{adv}(t) = A_{ij}(t) - \epsilon \times \text{sign}[x_i^* x_j(t)] \tag{S.7}$$

which corresponds to Eq. (4).

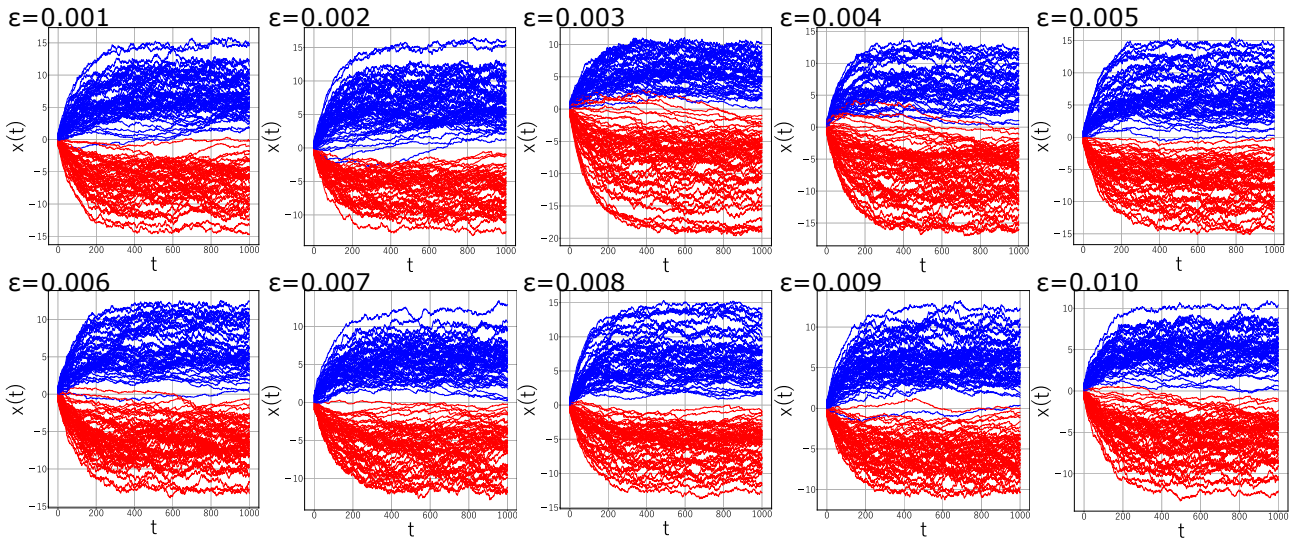

Figure S1: Opinion dynamics for varying values of ϵ within the range $\epsilon \in [0.001, 0.01]$. Each plot shows how the opinions $x(t)$ of agents evolve over time for a given value of ϵ. As ϵ increases, it weakens the connections between agents with similar opinions and strengthens the connections between those with opposing opinions, driving $\langle |x| \rangle$ closer to 0. However, for small values of ϵ (e.g., $\epsilon = 0.003, 0.004$), some agents overshoot, switching from positive to negative opinions (or vice versa), destabilizing the system and temporarily increasing $\langle |x| \rangle$.