The dynamics of opinion expression

Felix Gaisbauer, Eckehard Olbrich, and Sven Banisch

1 Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103 Leipzig, Germany

Abstract

We introduce a model of public opinion expression. Two groups of agents with different opinion on an issue interact with each other, changing the willingness to express their opinion according to whether they perceive themselves as part of the majority or minority opinion. We formulate the model as a multi-group majority game and investigate the Nash equilibria. We also provide a dynamical systems perspective: Using the reinforcement learning algorithm of Q-learning, we reduce the N-agent system in a mean-field approach to two dimensions which represent the two opinion groups. This two-dimensional system is analyzed in a comprehensive bifurcation analysis of its parameters. The model identifies structural conditions for public opinion predominance of different groups. We also show under which circumstances a minority can dominate public discourse. This offers direct connection to a central opinion expression theory of the social sciences, the spiral of silence.

1 Introduction

Fundamental to models of opinion dynamics is the assumption that people's opinions are, in some way or another, influenced by the opinion of their peers. The amount of models of opinion change in social systems is vast (see [2, 3, 4] for reviews). While it is a plausible assumption that people who express their opinion about an issue are sensitive to approval and disapproval, the effect of the interaction with others is not necessarily restricted to the opinion itself. It might also affect one's willingness of opinion expression: While one still thinks the others are wrong, one might not want to convey this publicly any more.

In comparison, this way of looking at public discourse has remained, from a modelling perspective, rather unexplored, albeit being meaningful in many situations: People usually do not change their opinions on a ‘day-to-day’ basis, but mostly within relatively long time scales, and once a population has become (bi-)polarized, public debate revolves mainly around which arguments or opinions are dominating the public discourse. We will hence, in this paper, focus on a model of the expression of opinions, and not the change in, opinions. An extensive theory of public opinion expression has already been developed 50 years ago, with Elisabeth Noelle-Neumann’s influential ‘spiral of silence’. Roughly speaking, Noelle-Neumann sees the fear of isolation as an essential drive for how humans publicly behave. Especially with respect to morally charged topics, individuals constantly and mostly sub-consciously monitor the ‘opinion landscape’ around them (they possess a “quasi-statistical sense” and might refrain from expressing their opinion if they believe to be part of the minority. On the other hand, a belief to hold the majority position might encourage them to express their view. Since each individual’s willingness or unwillingness to express her opinion influences how others perceive the opinion landscape, whose evaluation might then change accordingly, a dynamic development (what Noelle-Neumann called a ‘spiralling process’) follows in which the seemingly dominant opinion fraction becomes more and more vocal and the perceived minority fraction becomes more and more silent. Noelle-Neumann’s spiral of silence is particularly interesting for...
mathematical modelling since it links a micro mechanism with a dynamical development at the macro
level.

While there are contributions modelling opinion expression and specifically the spiral of silence, they
are either in large parts simulative \cite{6, 7, 8, 9, 10} or directed towards the effect of specific circumstances
on the spiral of silence (mass media \cite{11}, social bots \cite{12}, the long-time effect of charismatic agents
\cite{10}). Granovetter and Soong \cite{12}, and subsequently Krassa \cite{13}, employ a threshold model of opinion
expression which only applies to cases in which a certain opinion is already suppressed. We aim here
towards a more general, structural understanding of the dynamics of opinion expression.

We develop a model which employs an account of social influence similar to \cite{14}, where a model of
opinion dynamics was developed in which the opinion adjustment of an agent depends solely on the
social feedback she receives if she expresses her opinion. In the present approach, the effect of social
interaction is directed towards the willingness of or incentive for individuals to publicly express their
opinion. This presumably happens on much shorter time scales than opinion change. We investigate the
structural conditions that promote or hinder opinion expression of the different opinion groups. This is
firstly done from a game-theoretic angle, and subsequently combined with reinforcement learning in the
form of Q-learning. \cite{15} This allows us to perform a two-dimensional approximation of the system by a
mean-field approximation for the expected reward of the two opinion groups.

In the following, we will first describe the baseline social structure and the two central structural
parameters, $\gamma$ and $\delta$, of the model. In section 3 we represent the model as a multi-group majority game
on the agent network, and investigate its Nash equilibria with respect to $\gamma$ and $\delta$. Section 4 introduces
Q-learning and a subsequent two-dimensional approximation of the dynamical system. In section 5 we
perform a bifurcation analysis for the different parameters involved. We conclude with a discussion of
the results and an outlook in section 6.

2 Social-structural setting

For simplicity, we assume that there are two groups of individuals holding two different opinions on an
issue\footnote{The binary case is the generic one for polarized situations. In general, it may be a realistic assumption for some—such as referenda with only two choices—and an unrealistic one for other publicly discussed issues.}. The opinion of an agent $i$, $o_i$, is either 1 or 2, depending on the group she belongs to. The
group of agents holding opinion 1 is referred to as $G_1$, the group of agents holding opinion 2 as $G_2$. According
to their opinion, the connections between agents are described by weighted blocks according to the adjacency matrix $A$,

\begin{equation}
A = \begin{pmatrix}
q_{11} & q_{12} \\
q_{12} & q_{22}
\end{pmatrix}.
\end{equation}

Opinion group $G_1$ has size $N_1$ and opinion group $G_2$ has size $N_2$. The weight of an edge between any
two agents of community 1 is $q_{11}$, and analogously $q_{22}$ for the second community. Cross-edge weights
are given by $q_{12}$ and $q_{21}$. We assume an undirected network, hence $q_{12} = q_{21}$,

and the weights $q_{11}, q_{12}, q_{21} \in [0, 1]$.

The weights of the edges can be interpreted as the intensity or strength of the connections: The
smaller the weight, the less strong agents perceive the presence of each other\footnote{Alternatively, one could say that it is less likely that two agents encounter each other. $A$ is then interpreted as a stochastic block matrix and the weights as probabilities of there being an edge between agents.}. We can then give a formula for the perceived fraction of agents in the same group as an agent by

\begin{equation}
f_{11} = \frac{(N_1 - 1)q_{11}}{(N_1 - 1)q_{11} + N_2 q_{12}}
\end{equation}

for agents of opinion group $G_1$ and

\begin{equation}
f_{22} = \frac{(N_2 - 1)q_{22}}{(N_2 - 1)q_{22} + N_1 q_{12}},
\end{equation}

where $N_1$ and $N_2$ represent the size of the opinion groups $G_1$ and $G_2$, respectively.
for agents of opinion group $G_2$. The perceived fractions of neighbors belonging to the other opinion group are consequently

$$f_{12} = \frac{N_2 q_{12}}{(N_1 - 1) q_{11} + N_2 q_{12}},$$

(4)

and

$$f_{21} = \frac{N_1 q_{12}}{(N_2 - 1) q_{22} + N_1 q_{12}}.$$  

(5)

We now introduce the two central structural parameters, $\gamma$ and $\delta$. They are the ratios of the weighted in-group to the out-group connections for each opinion group and given by

$$\gamma = \frac{N_1 - 1}{N_2} \cdot \frac{q_{11}}{q_{12}},$$

(6)

and

$$\delta = \frac{N_2 - 1}{N_1} \cdot \frac{q_{22}}{q_{21}}.$$  

(7)

$\gamma > 1$ or $\delta > 1$ means that the agents of one opinion group are more strongly connected (under consideration of both the group sizes and the weights) to agents of the same than of the other opinion, while $\gamma < 1$ or $\delta < 1$ indicates that agents of the opinion group are more strongly connected to agents holding a different opinion. In the following, if we say that an opinion group is ‘internally well-connected,’ we mean that the structural parameter of the group is at least bigger than 1. With $\gamma$ and $\delta$, the above fractions (2), (3), (4) and (5) can be simplified to

$$f_{11} = \frac{\gamma}{\gamma + 1}, \quad f_{12} = \frac{1}{\gamma + 1},$$

(8)

$$f_{22} = \frac{\delta}{\delta + 1}, \quad f_{21} = \frac{1}{\delta + 1}.$$  

(9)

Alternatively, we can interpret the fractions $f_{11}$, $f_{22}$, $f_{12}$ and $f_{21}$ as the probabilities of interaction between agents of the different groups. This interpretation will be made use of in chapter 4.

### 3 A silence game

We formalize the sensitivity of individuals towards their opinion landscape and their fear of isolation from a game-theoretic perspective. We formulate a majority game [16] taking place in the social structure of chapter 2. The opinions of the agents are fixed according to their group affiliation and do not change. Agents act according to the perceived opinion proportions in their social environment. They can either choose opinion expression or silence as their action. If it appears to them that they are part of a minority opinion, they become silent. If they think that they are part of the majority, they express their opinion. But only the expressive agents shape the subjective impression of the opinion landscape of each individual. They do not count silent ones—after all, silence means that the individual’s opinion is not public.

Moreover, we introduce as an additional model assumption that opinion expression does not come for free. It is ‘costly’ to express one’s opinion, which is accounted for by a constant cost $c$. This constant might make more than a simple (perceived) opinion majority necessary for an agent to also have an incentive to express her opinion.

The ordinal preferences of an individual $i$ over the actions $e$ (for opinion expression) and $s$ (for silence) are given as follows: The payoff or utility for expressing her opinion, $U_i(e, a_{-i})$ (given the actions of all others $a_{-i}$) is bigger than the one for silence, formally

$$U_i(e, a_{-i}) > U_i(s, a_{-i}),$$

Multi-group games are considered e.g. in [17, 18].

We may think of the effort of typing a reply to someone in social media, or the effort of joining a demonstration for or against some issue.
if more agents in her perceived social environment speak out who share her opinion. We can state this condition formally by the inequality

\[
\frac{\sum_{j \neq i} q_{11} a_j}{(N_1 - 1)q_{11} + N_2 q_{12}} > \frac{\sum_{j \neq i} q_{12} a_j}{(N_1 - 1)q_{11} + N_2 q_{12}} + c \tag{10}
\]

if \( i \) is part of opinion group \( G_1 \) and

\[
\frac{\sum_{j \neq i} q_{22} a_j}{(N_2 - 1)q_{22} + N_1 q_{12}} > \frac{\sum_{j \neq i} q_{12} a_j}{(N_2 - 1)q_{22} + N_1 q_{12}} + c \tag{11}
\]

for \( i \) being part of \( G_2 \). Both the right- and the left-hand side are normalized with respect to the overall weighted connections of agent \( i \). Here, the actions \( a_j \) are given by \( a_j = 1 \) for expression and \( a_j = 0 \) for silence. If the two sides of the inequality are equal, the individual is indifferent in her preference over the actions.

A strategy profile is called a Nash equilibrium (NE) if no individual \( i \) can increase her expected reward by unilaterally deviating from the equilibrium. In our system, the equilibrium condition is met if there is a strategy profile for which each individual that expresses herself has \((10)\) or \((11)\) (depending on the opinion group of the agent) satisfied, and if for each individual that is silent, the corresponding inequality is not fulfilled.

It is already visible in \((10)\) and \((11)\) that apart from the fact that an individual does account for her own expressed opinion in the inequality (\( i \neq j \) in the sum on the left-hand side), the rest of the contributions in the inequalities are the same for all agents of one opinion group. It is also visible that if \((10)\) or \((11)\) is satisfied for an agent \( i \) that expresses herself, it must be satisfied for all silent individuals of her group as well: They ‘see’ one more agent expressing their opinion than \( i \), since \( i \) does not account for herself in her evaluation of her environment. Thus, \( i \) loses one summand of \( q_{11} \cdot 1 \) or \( q_{22} \cdot 1 \) in her inequality. On the other hand, if the inequality is not fulfilled for a silent agent of one group, it can also not be fulfilled for an expressive one. Therefore, in a pure-strategy equilibrium, all agents of one opinion group must choose the same action.

This simplifies the inequalities above. If all agents of an opinion group act the same, \((10)\) and \((11)\) can be expressed in terms of the structural parameters \( \gamma \) and \( \delta \). Four pure-strategy NEs might be possible, depending on \( \gamma \) and \( \delta \). Both groups can be silent, or only one of them, but not the other, or none:

- If both groups express their opinion (we call this state \((e, e)\); the first entry stands for the action of \( G_1 \), the second for the action of \( G_2 \)), the following conditions must be satisfied to make this a NE:

\[
\frac{(N_1 - 1)q_{11} - N_2 q_{12}}{(N_1 - 1)q_{11} + N_2 q_{12}} - c = \frac{\gamma - 1}{\gamma + 1} - c > 0 \tag{12}
\]

\[
\frac{(N_2 - 1)q_{22} - N_1 q_{12}}{(N_2 - 1)q_{22} + N_1 q_{12}} - c = \frac{\delta - 1}{\delta + 1} - c > 0. \tag{13}
\]

- \((e, s)\) is a NE if

\[
\frac{(N_1 - 1)q_{11}}{(N_1 - 1)q_{11} + N_2 q_{12}} - c = \frac{\gamma}{\gamma + 1} - c > 0, \tag{14}
\]

\[
\frac{N_1 q_{12}}{(N_2 - 1)q_{22} + N_1 q_{12}} - c = - \frac{1}{\delta + 1} - c < 0. \tag{15}
\]

- \((s, e)\) is a NE if

\[
\frac{N_2 q_{12}}{(N_1 - 1)q_{11} + N_2 q_{12}} - c = - \frac{1}{\gamma + 1} - c < 0, \tag{16}
\]

\[
\frac{(N_2 - 1)q_{22}}{(N_2 - 1)q_{22} + N_1 q_{12}} - c = \frac{\delta}{\delta + 1} - c > 0. \tag{17}
\]

- \((s, s)\) is a NE if

\[
-c < 0. \tag{18}
\]

\footnote{We use equation \((6)\) in the equivalence.}
The different existence regimes of the pure-strategy NEs are given in Figure 1. If γ and δ are both smaller than \( \frac{c}{1-c} \), then even if all group members express their opinion and the other opinion group is silent, it is too costly (compared to the amount of connections to agents of the own opinion group) to express one’s opinion and the only NE is the one in which all individuals are silent. If γ or δ or both are bigger than \( \frac{c}{1-c} \), but smaller than \( \frac{c+1}{1-c} \), either both opinion groups are silent or one of the groups expresses themselves, but not both: The strength of internal connections of each group are not sufficient to account for the negative influence of the other, expressive group. Not both (12) and (13) can be satisfied. Hence, this structural regime only allows public opinion predominance of one group (or complete silence). If γ and δ are both bigger than \( \frac{c+1}{1-c} \), it is possible that both opinion groups express their opinion publicly at the same time. Then, the positive influence of the in-group members still dominates, even if all out-group members are expressive as well. Hence, also (12) and (13) are satisfied.

Obviously, there are also mixed-strategy Nash equilibria. Suppose the situation is as follows: The agents of each group mix their actions uniformly such that each agent is exactly indifferent between expressing herself or staying silent. Then, no one has the incentive to change their action, and we therefore have a NE. This equilibrium is, nevertheless, only metastable in the sense that it only takes one agent to increase (or decrease) her expression probability in order to make it favourable for all other agents of one opinion group to express themselves (or become silent).

γ and δ do not only depend on the number of agents holding one or another opinion. They are also influenced by the internal connection weights of agents of one opinion group. Hence, a well-connected minority group can dominate public discourse if the corresponding structural parameter is above the threshold of \( \frac{c}{1-c} \). While the regimes of different NEs in Figure 1 are displayed correctly, it might give the impression that γ or δ are parameters that can be tuned by simply increasing the weight of a connection between two agents of the same group, that is, \( q_{11} \) or \( q_{22} \). That is not the case. Some numerical minorities cannot be balanced by increasing internal connections since \( q_{11} \) and \( q_{22} \) are bounded by 1. If there are too few agents in one opinion group, even setting \( q_{11} \) or \( q_{22} \) to 1 will not be elevate γ or δ above a certain threshold. This is made visible in Figure 2. The figure shows the different existence regimes of the NEs for different combinations of internal connection weights \( q_{11} \) and \( q_{22} \) and partitions of a total of \( N = 100 \) agents between groups \( G_1 \) and \( G_2 \), \( q_{12} \) and \( c \) are fixed. Each point in the plot stands for a combination of the number of agents in opinion group \( G_1 \), \( N_1 \), and the in-group connection weights \( q_{11} \), out of which one can compute the value of γ. The lines of constant γ are plotted in red. Since the overall number of agents \( N = 100 \) is fixed, \( N_2 \) is not independent and determined by the choice of \( N_1 \).
Figure 2: The constant $\gamma$- and $\delta$-curves for $N = 100$ agents. They are plotted with respect to $N_1$, $N_2 = N - N_1$, and $q_{11} = q_{22}$. Each blue curve stands for a combination of the number of opinion group members $N_1$ and internal connection weights $q_{11}$ that yields a constant value of the structural parameters $\gamma$, each red one for a combination of $N_2$ and $q_{22}$ that produces constant $\delta$. The color-coding for the different Nash equilibrium regions is analogous to Figure 1. It is visible that the numerical minority of an opinion group cannot always be compensated by increasing $q_{11}$ (or $q_{22}$), the weight of a connection between two agents of the same opinion group. Moreover, the fixed $\gamma$- and $\delta$-curves are symmetric with respect to $N_1 = N_2 = 50$, where they intersect. (For better readability, the dashed $\delta$-curves have not been labelled. They correspond to their $\gamma$-counterparts.)

by $N - N_1$. If we just assume that $q_{22} = q_{11}$, each point in the plot at the same time represents also a combination of the relevant parameters of opinion group $G_2$ out of which one can compute $\delta$. Curves of constant $\delta$ are the blue lines and symmetrical to the $\gamma$-curves with respect to $N_1 = 50$.

A vertical line in the plot, e.g. at $N_1 = 20$, can be interpreted as follows: Each constant $\gamma$ or $\delta$ value that it intersects on its way to $q_{11} = q_{22} = 1$ is reachable for this partition of agents in the two groups if $q_{11}$ and $q_{22}$ are tuned accordingly. But if there is no intersection for a specific $\gamma$ or $\delta$, then even if the internal connection weights are maximized, the structural strength of the respective group cannot reach that value due to their limited group size. For $N_1 = 20$, a state in which both opinion groups are expressing themselves (the upper right, green area in Figure 1) cannot be reached since opinion group $G_1$ has too few agents to produce a $\gamma$ high enough to satisfy (12). In general, there are numerical thresholds (dependent on the costs $c$, the cross-group connection weight $q_{12}$ and the overall number of agents $N$) below which reaching a state in which both group express themselves or reaching a state in which the own group becomes dominant becomes impossible from a game-theoretic perspective.

4 Q-learning and a dynamical systems perspective

While we are able to determine the Nash equilibria of the system, the game-theoretical point of view does not answer questions of equilibrium selection or the effects of bounded rationality. In this section, we will introduce a dynamical systems perspective to approach those questions.

We posit a simple interaction mechanism between the agents on the network of section 2. It is given as follows: If an agent expresses her opinion, she will be paired with a random neighbor. Now, the fractions $f_{11}, f_{12}, f_{21},$ and $f_{22}$ correspond to the probability of meeting a neighbor of a certain opinion group given the own opinion group of an agent. The neighbor then gives (if she also is in an expressive state) social feedback to the agent, either agreement or disagreement, which will shape the agent’s impression of her opinion environment. Put in an algorithmic way:

1. A random agent is selected.
2. If willing to speak out, the agent expresses her opinion to a random neighbor at cost $c$.
3. If the neighbor is also willing to speak out, she gives feedback on the agent’s opinion.
4. According to the feedback, the agent will become more/less willing to speak out.
As in [15], we will describe the development of the system as reinforcement learning dynamics, more specifically, as dynamics induced by Q-learning. In Q-learning, the reinforcement mechanism that updates the agent’s willingness to express her opinion is given by
\[ Q_{i}^{t+1} = (1 - \alpha)Q_{i}^{t} + \alpha r_{i}^{t}, \]  
where \( r_{i}^{t} \) gives the reward for agent \( i \) at time step \( t \) upon expression
\[ r_{i}^{t} = \begin{cases} 
-c & \text{for random neighbor being silent,} \\
-c - 1 & \text{for disagreeing random neighbor,} \\
-c + 1 & \text{for agreeing random neighbor.} 
\end{cases} \]  
The Q function should converge with time to the expected reward.\(^8\) The probability of expression is a function of the value of \( Q_{i} \). We assume here a Boltzmann action selection mechanism, i.e. the probability of expression of agent \( i \) is given by
\[ p_{i}^{t} = \frac{1}{1 + e^{-\beta Q_{i}^{t}}}, \]  
the probability of staying silent by \( 1 - p_{i}^{t} \). If \( \beta = 0 \), the action choice of the agent is completely independent of the \( Q \)-values and randomized. For increasing \( \beta \), the agent becomes more sensitive in her action selection towards her current evaluation of her local opinion environment. Then, a positive \( Q \)-value indicates that it is more likely for her to express herself than not, while a negative one indicates the opposite. If \( \beta \to \infty \), the probabilities of the actions become deterministic.

The expected reward for agent \( i \) upon opinion expression is given by either (if \( i \) belongs to opinion group \( G_{1} \))
\[ E_{p}[r_{i}^{t}] = -c + f_{11} \frac{1}{N_{1}} - 1 \sum_{j \in G_{1} \setminus i} \frac{1}{1 + e^{-\beta Q_{j}^{t}}} - \\
f_{12} \frac{1}{N_{2}} \sum_{j \in G_{2}} \frac{1}{1 + e^{-\beta Q_{j}^{t}}}, \]  
or (if \( i \) belongs to opinion group \( G_{2} \))
\[ E_{p}[r_{i}^{t}] = -c + f_{21} \frac{1}{N_{1}} - 1 \sum_{j \in G_{1} \setminus i} \frac{1}{1 + e^{-\beta Q_{j}^{t}}} - \\
f_{22} \frac{1}{N_{2}} \sum_{j \in G_{2}} \frac{1}{1 + e^{-\beta Q_{j}^{t}}}. \]  
We follow [15], where Q-learning in two-player two-action games is investigated, and take the continuous-time limit of the Q-learning equation [19]. We divide time there into intervals \( \delta t \). We replace \( t + 1 \) with \( t + \delta t \) and \( \alpha \) with \( \alpha' \delta t \). This yields
\[ Q_{i}(t + \delta t) - Q_{i}(t) = \alpha' \delta t (r_{i}(t) - Q_{i}(t)) \]  
and hence
\[ \dot{Q}_{i} = \alpha'(r_{i}(t) - Q_{i}(t)). \]  
If we now look at the time development of the \( Q \)-values, the difference of the largest and the lowest \( Q \)-value of an opinion group decays at least exponentially in expectation (see the Appendix for the estimation):
\[ \frac{d}{dt}(Q_{i \in G_{1}}^{\max} - Q_{i \in G_{1}}^{\min}) \leq -\alpha'(Q_{i \in G_{1}}^{\max} - Q_{i \in G_{1}}^{\min}), \]  
\[ \frac{d}{dt}(Q_{i \in G_{2}}^{\max} - Q_{i \in G_{2}}^{\min}) \leq -\alpha'(Q_{i \in G_{2}}^{\max} - Q_{i \in G_{2}}^{\min}). \]  
That is, the \( Q \)-values of the agents of one group are expected to converge over time. We now employ a mean-field approximation for the expected reward of the two opinion groups: We introduce the average \( Q \)-values for each opinion group
\[ Q_{1}(t) = \frac{1}{N_{1}} \sum_{i \in G_{1}} Q_{i}(t), \quad Q_{2}(t) = \frac{1}{N_{2}} \sum_{i \in G_{1}} Q_{i}(t). \]  
\(^{19}\) describes Q learning for myopic agents, i.e. with discount factor 0.
This means that we will not distinguish any more between the agents of the respective opinion groups, but will assign them the average of their group’s $Q$-value. This simplification will have an effect on the probability of opinion expression for the individuals. Instead of averaging over each groups probability of expression, we simply insert the averaged $Q$-values into the equation:

$$
\frac{1}{N_1} \sum_{j \in G_1} \frac{1}{1 + e^{-\beta Q_j(t)}} \rightarrow \frac{1}{1 + e^{-\beta Q_1(t)}} = p_1(t),
$$

(26)

$$
\frac{1}{N_2} \sum_{j \in G_2} \frac{1}{1 + e^{-\beta Q_j(t)}} \rightarrow \frac{1}{1 + e^{-\beta Q_2(t)}} = p_2(t).
$$

(27)

The expected reward for the respective opinion groups can now be represented by the equations:

$$
E_p[r_1(t)] = -c + \frac{\gamma}{\gamma + 1} p_1(t) - \frac{1}{\gamma + 1} p_2(t),
$$

(28)

$$
E_p[r_2(t)] = -c + \frac{\delta}{\delta + 1} p_2(t) - \frac{1}{\delta + 1} p_1(t),
$$

(29)

where the probabilities of expression for each group are simply given by $p_1(t)$ and $p_2(t)$, and it is not distinguished any more between the individuals.

We can therefore write our two-dimensional formulation as follows:

$$
\dot{Q}_1 = \alpha'(-c + \frac{\gamma}{\gamma + 1} p_1(t) - \frac{1}{\gamma + 1} p_2(t) - Q_1(t)),
$$

(30)

$$
\dot{Q}_2 = \alpha'(-c + \frac{\delta}{\delta + 1} p_2(t) - \frac{1}{\delta + 1} p_1(t) - Q_2(t)).
$$

(31)

According to equations (30) and (31), we can produce a phase portrait of the system including its trajectories and fixed points for given exploration rate $\beta$, structural parameters $\gamma$ and $\delta$, and costs of expression $c$. An example of the development of such phase portraits with changes in $\gamma$ and $\delta$ is given in Figure 3.

There, it is visible that all stable fixed points of the system include basins of attraction, that is, regimes of values of $Q_1$ and $Q_2$ for which the system is expected to end up in those fixed points. The basins of attraction in the two-dimensional approximation correspond exactly to those of the stochastic $N$-agent system in the limit $\alpha \rightarrow 0$. For larger $\alpha$, both fixed points and basins of attraction do not necessarily correspond to the two-dimensional approximation. We show averages over simulation runs for different values of $\alpha$ in Figure 4.

5 Bifurcation and stability analysis

In order to find the fixed points of $Q_1$ and $Q_2$, we set (30) and (31) 0, solve (30) for $Q_2$ and insert it into (31), which yields:

$$
Q_2 = -\frac{1}{\beta} \ln(\frac{\gamma}{1 + e^{-\beta Q_1}} - (\gamma + 1)(Q_1 + c) - 1)
$$

(32)

$$
0 = \frac{\delta}{\delta + 1}(\frac{\gamma}{1 + e^{-\beta Q_1}} - (\gamma + 1)(Q_1 + c)) - c - \frac{1}{\delta + 1} + \frac{1}{\delta + 1} e^{-\beta Q_1} + \frac{1}{\beta} \ln(\frac{\gamma}{1 + e^{-\beta Q_1}} - (\gamma + 1)(Q_1 + c) - 1)
$$

(33)

Equation (33) now gives us the $Q_1$-value of the fixed points of the system, with which we can calculate the corresponding $Q_2$-value by equation (32). In essence, the fixed points depend on four parameters: $\beta$, $\gamma$, $\delta$, and $c$. We will carry out a bifurcation analysis of these parameters in the following.

After having solved equations (33) and (32) for $Q_1$ and $Q_2$, we can assess the stability of the respective fixed points by calculating the eigenvalues of their Jacobian; two negative (real parts of the) eigenvalues indicate a stable attractor. In the following, we analyze the bifurcation structure of the system depending on the different types of parameters in the system.

\[ f_{11}, f_{12}, f_{21}, \text{and} \ f_{22} \text{have been replaced according to equations} (\text{8}) \text{and} (\text{9}) \text{with} \ \frac{\gamma}{\gamma + 1}, \ \frac{1}{\gamma + 1}, \ \frac{\delta}{\delta + 1}, \text{and} \ \frac{1}{\delta + 1}. \]
Figure 3: The development of the phase portrait of the two-dimensional system in the parameter space of $\gamma$ and $\delta$. We have $c = 0.1$, $\beta = 10$, and structural parameters $\gamma = \delta = 0.1$ (bottom left), $\gamma = \delta = 1$ (bottom right), and $\gamma = \delta = 3$ (top right). Different fixed points arise for these parameter constellations.
Figure 4: The trajectories of the $Q$-values in simulations, averaged over 50 runs with $N \cdot 10^5$ steps, for different values of $\alpha$ with a starting point close to the border (light red) of the two basins of attraction of the two fixed points (starting $Q$-values were $Q_{i \in G_1} = 0$, $Q_{i \in G_2} = -0.25$). There were $N = 200$ agents, 100 of each opinion group, and $c = 0.1$, $q_{11} = 0.04$, $q_{12} = 0.05$, and $q_{22} = 0.15$. A relatively big $\alpha = 0.1$ makes the trajectory leave the lower right basin of attraction of the two-dimensional system. Due to the high $\alpha$, the fixed point of the other basin of attraction is also missed by some margin. The lower $\alpha$, the closer the trajectories get to the fixed point and the more probable it is that they will stay in the basin predicted by the two-dimensional approximation. For $\alpha = 0.01$ and $\alpha = 0.001$, the trajectories run towards the predicted fixed point.

5.1 Structural power

The parameter $\gamma$ describes the ratio of internal versus external connectedness of one opinion group. $\gamma > 1$ means that on average each member of one opinion group is connected more strongly to agents of the own than of the other opinion group. As is visible in Figure 5 for small $\gamma$ (< 0.5), given $\beta = 10$, $\delta = 2.36$ (that is, a quite well-connected opposite opinion group) and $c = 0.1$, there is only one (stable) fixed point with negative $Q_1$-value and positive $Q_2$. While $\gamma$ grows, a saddle-node bifurcation occurs such that one stable and one unstable fixed point appear for positive $Q_1$ and negative $Q_2$. Another saddle-node bifurcation occurs at around $\gamma = 2$; and for $\gamma > 4.2$, the low-$Q_1$ fixed points disappear in another saddle-node.

How can this be interpreted? In essence, an opinion community that is not well-connected internally ($\gamma < 0.5$) will be driven into silence (a $Q$-value much lower than 0) by the opposite opinion group that is internally more cohesive. With increasing $\gamma$, that is, increasing internal connectedness, other fixed points, in which the former group is expressive, appear. To be precise, the $Q$-values here are only indicative of probabilities of opinion expression according to the Boltzmann action selection which depends on $Q$. If $Q$ is smaller than 0, the probability of expression is smaller than the probability of staying silent. In the following, if we say that one opinion group is expressive, we mean that they have a $Q$-value bigger than 0 which makes their probability of expression higher than that of silence. With a further increase of $\gamma$, the stable fixed point for which only the opposite opinion group is expressive disappears and we remain with three fixed points (the middle one unstable), for which either the first opinion group is ‘loud’ alone or both groups express their opinions. Hence, $G_1$ is now too cohesive to be driven into silence by the other group. A lower $\delta$-value (e.g. $\delta = 1.6$) leads to a reduction in available fixed points (Figure 6) such that only two saddle-node bifurcations occur and at high $\gamma$ only one fixed point remains in which the first opinion group is expressive.

---

10Everything stated in this paragraph applies equivalently to $\delta$, which is just the parameter for the ratio of internal versus external connectedness of the other group.
Figure 5: The development of the fixed points with $\gamma$ given $\beta$, relatively high inner connectedness $\delta$ of $G_2$, and $c$. The colors of the curves in the two plots correspond to each other and indicate the different fixed point pairs of $Q_1$ and $Q_2$. A dashed line indicates an unstable fixed point, a continuous one a stable fixed point. It is visible in the plots that a poorly connected opinion group $G_1$ ($\delta < 0.5$) will be driven into silence by the other group (beige curve). With increasing in-group connectivity, fixed points arise for which $G_1$ expresses their opinion in two saddle-node bifurcations (red for an an $(e,s)$-equilibrium and blue for $(e,e)$). For $\gamma > 4.5$, $G_1$ is so well-connected that the equilibrium disappears in which the group is silent.

Figure 6: The development of $Q_1$- and $Q_2$-fixed points with $\gamma$ given $\beta$, moderate $\delta$ and $c$. For $\gamma < 0.4$, only group $G_2$ is expressive. A second fixed point arises for higher $\gamma$ in which $G_1$ is predominating public discourse. There is no fixed point in which both groups are expressive.
5.2 Costs

The costs for opinion expression have a profound impact on the fixed points of the system. If opinion expression is very ‘expensive,’ (in Figure 7: $c > 0.4$), there is only one fixed point in the system for which both opinion groups stay silent. For decreasing costs, two pairs of fixed points arise in a saddle-node bifurcation. Each of the pairs corresponds to a situation in which one opinion group is expressive, while the other is silent (in Figure 7, we have identical values for $\gamma$ and $\delta$). The fixed point in which both opinion groups are silent becomes unstable with decreasing $c$ in a pitchfork bifurcation. Below $c = 0.1$, another pitchfork bifurcation arises for which the stable fixed point now corresponds to a state in which both groups are expressing their opinion. Costs can also be negative: Then, the individuals are intrinsically motivated to speak out. For sufficiently negative costs (in the case of Figure 7: $c < -0.5$), only one fixed point exists: Everyone has an incentive to speak out, at least for internally well-connected opinion groups. The fixed points for which only one of the groups is expressive disappear in two saddle-nodes.

5.3 Asymmetric costs

It is quite an artificial assumption that costs for opinion expression are equal for both opinion groups. Internal motivation for a cause, for example, can be an incentive to speak out and might even be indicated by negative costs (that is, an urge to express one’s opinion). If one group has a stronger motivational disposition than the other to speak out, we might face a situation in which $c_1 \neq c_2$.

The bifurcation in Figure 8 illustrates the effect that different motivational dispositions in the populations exhibit on public discourse. In Figure 8, a bifurcation over $c_1$ is shown. Very low motivation (that is, high $c_1$) for opinion expression in opinion group $G_1$ yield one equilibrium in which opinion group $G_1$ is the only expressive one.

11Ideals such as e.g. free speech might have such an effect: People then see it as their duty to voice their opinion, especially if it does not conform to the apparent majority.
1 is silent, while opinion group $G_2$ is expressive. With stronger motivation, two saddle-node bifurcations appear one after the other. The one appearing at $c_1 \approx 0.4$ includes a stable equilibrium in which group 1 dominates public discourse, and in the one appearing at $c_1 \approx 0.15$, both groups are expressive. When $c_1$ becomes smaller than 0, the equilibrium in which group $G_1$ is silent disappears in a saddle-node since they are too strongly motivated.

5.4 Exploration rate

The parameter $\beta$ determines how sensitive agents are in their actions towards the current evaluation of their expected reward. A high $\beta$-value indicates a choice of the agent similar to a best response to their current evaluation of the expected reward, while $\beta = 0$ means that each available action is chosen with equal probability.

As is visible in Figure 9, for very low $\beta$, there is only one fixed point available with a very low $Q$-value for both opinion groups. With increasing sensitivity ($\beta \approx 5$) towards the current perceived state of the opinion landscape, further fixed points arise in a supercritical pitchfork bifurcation, and then, at $\beta > 6$, another (now subcritical) pitchfork bifurcation arises, such that we arrive at three stable fixed points (one in which both groups are in an expressive mode, and one for opinion dominance for each group) and two unstable ones in-between.

6 Discussion and outlook

The spiral of silence and beyond

The present model provides a structural view on collective opinion expression. It reproduces the counterintuitive result postulated by Noelle-Neumann in her theory of the spiral of silence [1, 5], namely the possibility of the public dominance of a minority opinion. While the influence of mass media has been stressed in many publications concerning the spiral of silence, we show that no mass media is needed for this effect. Being an internally well-connected community alone can be enough to gain public dominance. This finding gains increasing traction in light of the advent of social media, which facilitated communication among like-minded people and decentralized information distribution. Apart from that, the present approach also provides conditions for the ‘overcoming’ of the spiral of silence, for which the numerical proportions do not necessarily have to change. The increase in internal cohesion of the different opinion groups can be sufficient. On the other hand, it is also shown that if the minority is too small, even maximum internal cohesion cannot heave the minority opinion into public predominance (see again Figure 2).

Perception biases

In [6], the effect of the ego-network size, that is, the (average) number connections of the agents, on the onset of the spiral of silence was investigated. It was concluded that an increase in network density makes silencing effects more probable. In our work, we show that more density might even have the opposite effect. It depends on where the additional connections are made: If new connections are guided by some homophily principle, such that the opinion blocks become more cohesive, the spiral of silence might even be overcome (see path (i) in Figure 9). We then arrive at a structure similar to the infamous ‘echo chambers’, in which only the voices affirming one’s own view are heard and the others are blocked out. If the additional connections are made between the opinion blocks, both $\gamma$
Figure 10: Illustration of the transitions between the game-theoretic equilibrium regions for (i) stronger internal cohesion of both opinion groups (‘echo chambers’), (ii) less internal cohesion of both (heterophilous connections), and (iii) stronger internal cohesion for only one opinion group (‘#metoo’).

and \( \delta \) decrease, which might make it more probable that the individuals have a more realistic picture of the overall opinion landscape. Then, the spiral effect is indeed more probable. But if the cross-group connections grow even further, both opinion groups misjudge their proportion to their own disadvantage, such that no group speaks out if there are costs associated to opinion expression (path (ii) in Figure 10). This is closely linked to more general accounts of perception biases [19]. Misjudgements are not necessarily caused by biased cognitive or motivational processes, but solely by the structure of the social contacts. The present model provides an instantiation thereof.

**Critical mass** Furthermore, the model links to studies dealing with tipping points in social systems and the necessary numerical allocations, depending on the network structures. This has e.g. been analyzed for social conventions [20]. If the social network of individuals is structured in opinion blocks, there is a hard numerical limit for the overcoming of a state in which one opinion is dominating publicly. For example, for a cross-opinion connection probability of 0.2 (as in Figure 2), the state in which both opinion camps are expressive cannot be reached if the minority makes up less than 22 per cent of the population.

**Limits and outlook** While we have stressed the generality of this work, we want to emphasize its limits as well: The homogeneous network structure of opinion blocks is not particularly realistic. Real social networks are rather heterogeneous, with well-connected and very active hubs and more ‘remote’ individuals. Nevertheless, weighted (or stochastic) blocks can serve as a baseline for mathematical accessibility.

Moreover, this work is concerned with one way of reacting on social feedback, namely, the change in willingness to express one’s opinion. Change in opinion is not included. As we have argued in the introduction, it is probable that these phenomena take place on different time scales. Also, the social environments prompting opinion change might be different from the ones in which opinion predominance is fought for. In demonstrations, if two opinion camps meet each other, the main objective is usually to assert opinion dominance and audibility, not information exchange or the need to convince each other. Hence, a combination of models of opinion change and opinion expression might be in order in a multi-layer network approach, in which opinion formation and the competition for public opinion predominance take place on possibly different network structures.

While there are plenty of studies on experimental evidence for the micromechanisms grounding the spiral of silence (see [21] for a review), we are also seeking a more systematic larger-scale view on collective phenomena of opinion expression, which are closely related to the parameters \( \gamma \) and \( \delta \) in the model. A very prominent example of emerging collective opinion expression online, for which this model provides an explanation, is the Twitter-hashtag ‘#metoo’ and the subsequent movement against sexual harassment and sexual assault: Women found a device (in this case, a hashtag) that allowed them to find and connect to people who had experienced the same, and also to people who supported them. And all of a sudden, it was easier for them to speak out (path (iii) in Figure 10). Measurements are an intricate task here:

\[ 12 \text{No matter if one finds this reasonable or not.} \]
The networks one constructs out of interactions between individuals are only the networks of interaction, that is, of only half of the actions one wants to observe. Silent individuals do usually not show up in such networks since they are not involved in an observable way. This must be kept in mind when dealing with any network representations of, especially, online interactions.

Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 732942 (www.ODYCCEUS.eu). We are grateful for the repeated discussion of the ideas described in the paper with Wolfram Barfuss, Sharwin Rezagholi, Florentin Münch, Armin Pournaki, and with the members of ODYCCEUS.

A Expected decrease of the difference in \( Q \)-values

We carry out the estimation for opinion group \( G_1 \), but the analogue holds for opinion group \( G_2 \). We can give an upper bound for the change in \( Q \)-value for the agent with the maximum \( Q \)-value of the group, \( Q_{i\in G_1}^{\text{max}} \), and a lower bound for the change in \( Q \)-value for the agent with the minimum \( Q \)-value of the group, \( Q_{i\in G_1}^{\text{min}} \), due to the monotonicity of the function \( \frac{1}{1+e^x} \):

\[
\dot{Q}_{i\in G_1}^{\text{max}} = \alpha' \left( \frac{\gamma}{\gamma + 1} N_1 - 1 \right) \sum_{j\in G_1, j\neq i} \frac{1}{1 + e^{-\beta Q_j}} - \frac{1}{1 + e^{-\beta Q_{i\in G_1}^{\text{max}}}} - Q_{i\in G_1}^{\text{max}} - c \leq \alpha' \left( \frac{\gamma}{\gamma + 1} N_1 \right) \sum_{j\in G_1, j\neq i} \frac{1}{1 + e^{-\beta Q_j}} + \frac{1}{1 + e^{-\beta Q_{i\in G_1}^{\text{max}}}} - \frac{1}{1 + e^{-\beta Q_{i\in G_1}^{\text{max}}}} - Q_{i\in G_1}^{\text{max}} - c,
\]

(34)

\[
\dot{Q}_{i\in G_1}^{\text{min}} = \alpha' \left( \frac{\gamma}{\gamma + 1} N_1 - 1 \right) \sum_{j\in G_1, j\neq i} \frac{1}{1 + e^{-\beta Q_j}} - \frac{1}{1 + e^{-\beta Q_{i\in G_1}^{\text{min}}}} - Q_{i\in G_1}^{\text{min}} - c \geq \alpha' \left( \frac{\gamma}{\gamma + 1} N_1 \right) \sum_{j\in G_1, j\neq i} \frac{1}{1 + e^{-\beta Q_j}} + \frac{1}{1 + e^{-\beta Q_{i\in G_1}^{\text{min}}}} - \frac{1}{1 + e^{-\beta Q_{i\in G_1}^{\text{min}}}} - Q_{i\in G_1}^{\text{min}} - c.
\]

(35)

If we now look at the change in time in the difference of \( Q_{i\in G_1}^{\text{max}} \) and \( Q_{i\in G_1}^{\text{min}} \), we can conclude by the above inequalities that the difference decreases at least exponentially in expectation by substracting the right hand-sides of (34) and (35).

\[
\frac{d}{dt}(Q_{i\in G_1}^{\text{max}} - Q_{i\in G_1}^{\text{min}}) \leq -\alpha'(Q_{i\in G_1}^{\text{max}} - Q_{i\in G_1}^{\text{min}}).
\]

(36)

The analogue holds for opinion group \( G_2 \):

\[
\frac{d}{dt}(Q_{i\in G_2}^{\text{max}} - Q_{i\in G_2}^{\text{min}}) \leq -\alpha'(Q_{i\in G_2}^{\text{max}} - Q_{i\in G_2}^{\text{min}}).
\]

(37)

References

[1] E. Noelle-Neumann and T. Petersen, “The spiral of silence and the social nature of man,” in Handbook of political communication research, pp. 357–374, Routledge, 2004.

[2] J. Lorenz, “Continuous opinion dynamics under bounded confidence: A survey,” International Journal of Modern Physics C, vol. 18, pp. 1819–1838, Dec. 2007.

[3] C. Castellano, S. Fortunato, and V. Loreto, “Statistical physics of social dynamics,” Reviews of Modern Physics, vol. 81, pp. 591–646, May 2009.

[4] A. Beverungen and J. W. van der Linden, “The role of interaction in network structures,” European Journal of Information Systems, vol. 13, pp. 190–198, 2004.

[5] G. M. Schwartz and J. P. Hoffman, “Opinion dynamics and social networks,” Social Networks, vol. 27, pp. 133–151, 2005.
[4] A. Flache, M. Märs, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, and J. Lorenz, “Models of social influence: Towards the next frontiers,” *Journal of Artificial Societies and Social Simulation*, vol. 20, no. 4, p. 2, 2017.

[5] E. Noelle-Neumann, “The spiral of silence a theory of public opinion,” *Journal of Communication*, vol. 24, pp. 43–51, June 1974.

[6] D. Sohn and N. Geidner, “Collective Dynamics of the Spiral of Silence: The Role of Ego-Network Size,” *International Journal of Public Opinion Research*, vol. 28, pp. 25–45, 03 2015.

[7] B. Ross, L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, and S. Stieglitz, “Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks,” *European Journal of Information Systems*, vol. 28, no. 4, pp. 394–412, 2019.

[8] D. Takeuchi, G. Tanaka, R. Fujie, and H. Suzuki, “Public opinion formation with the spiral of silence on complex social networks,” *Nonlinear Theory and Its Applications, IEICE*, vol. 6, no. 1, pp. 15–25, 2015.

[9] W. Waldherr and M. Bachl, “Simulation gesellschaftlicher medienwirkungsprozesse am beispiel der schweigespirale,” in *Rezeption und Wirkung in zeitlicher Perspektive*, pp. 235–252, Nomos Verlagsgesellschaft mbH & Co. KG, 2011.

[10] P. Gawronski, M. Nawojczyk, and K. Kulakowski, “Opinion formation in an open system and the spiral of silence,” *Acta Physica Polonica A*, vol. 127, 2015.

[11] D. Sohn, “Spiral of silence in the social media era: A simulation approach to the interplay between social networks and mass media,” *Communication Research*, June 2019.

[12] M. Granovetter and R. Soong, “Threshold models of diversity: Chinese restaurants, residential segregation, and the spiral of silence,” *Sociological Methodology*, vol. 18, pp. 69–104, 1988.

[13] M. A. Krassa, “Social groups, selective perception, and behavioral contagion in public opinion,” *Social Networks*, vol. 10, pp. 109–136, June 1988.

[14] S. Banisch and E. Olbrich, “Opinion polarization by learning from social feedback,” *The Journal of Mathematical Sociology*, vol. 43, pp. 76–103, Oct. 2018.

[15] A. Kianercy and A. Galstyan, “Dynamics of boltzmann q learning in two-player two-action games,” *Physical Review E*, vol. 85, no. 4, p. 041145, 2012.

[16] M. O. Jackson and Y. Zenou, “Games on networks,” in *Handbook of game theory with economic applications*, vol. 4, pp. 95–163, Elsevier, 2015.

[17] P. R. Neary, *Multiple-group games*. PhD thesis, UC San Diego, 2011.

[18] P. R. Neary, “Competing conventions,” *Games and Economic Behavior*, vol. 76, no. 1, pp. 301–328, 2012.

[19] E. Lee, F. Karimi, C. Wagner, H.-H. Jo, M. Strohmaier, and M. Galesic, “Homophily and minority-group size explain perception biases in social networks,” *Nature human behaviour*, pp. 1–10, 2019.

[20] D. Centola, J. Becker, D. Brackbill, and A. Baronchelli, “Experimental evidence for tipping points in social convention,” *Science*, vol. 360, pp. 1116–1119, 06 2018.

[21] J. Matthes, J. Knoll, and C. von Sikorski, “The “spiral of silence” revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression,” *Communication Research*, vol. 45, no. 1, pp. 3–33, 2018.