

友好的・敵対的関係を考慮した意見分極化の緩和手法

中条 雅貴 | 東京大学大学院工学系研究科システム創成学専攻 助教

劉 庶 | 株式会社 JDSC データサイエンティスト

鳥海 不二夫 | 東京大学大学院工学系研究科システム創成学専攻 教授

要約

現代社会ではオンライン空間を通じて多様な情報に触れられる一方で、近い意見ばかりが強化され、異なる立場の人々との対立が深まることもある。このような意見の分極化は社会的分断につながりうる重要な問題である。本研究では信頼する相手には同調し、不信感をもつ相手には反発しやすいという人間関係の特徴に着目し、これを上手く調整することで分極化の進行を抑制することを目的とする。友好的関係と敵対的関係を表現できる符号付きネットワーク上の意見形成モデルを用い、個人の意見を直接変えるのではなく、誰の意見がどの程度届きやすいかをわずかに調整する介入の効果を検証した。実世界のネットワークデータを用いた数値シミュレーションの結果、無作為な調整よりも現在の意見状態と関係性を考慮した提案手法の方が集団全体を合意に近づけやすいことが示された。このことから、分極化の抑制には単に情報接触を増やすだけでなく、人々の関係性や意見状態を踏まえた働きかけが有効であることが示唆される。本研究はオンライン空間で生じる過度な対立を理解し、分極化を抑制するための情報環境設計につながる知見を与える。



中条 雅貴

2020年北陸先端科学技術大学院大学融合科学共同専攻 修士課程修了。2023年北陸先端科学技術大学院大学融合科学共同専攻 博士後期課程修了、博士（融合科学）。同年より東京大学大学院工学系研究科 特任助教、2026年より助教として従事。専門分野は複雑ネットワーク、計算社会科学。

1. はじめに

現代社会においてオンライン情報空間は、人々が日常的に情報を取得し、意見を形成し、他者と議論するための重要な社会的インフラとなっている。ソーシャルメディア、検索エンジン、ニュース配信サービス、推薦システムなどの情報流通基盤は、個人が膨大な情報へ容易にアクセスすることを可能にし、政治、経済、科学、健康、災害対応など、さまざまな領域における情報伝達と意思決定を支えている。このような環境は多様な知識や視点に触れる機会を広げるという大きな利点をもつ。一方で、利用者の関心や過去の行動に基づいて情報が選別されるため、特定の意見や価値観に偏った情報に継続的に接触しやすくなるという問題も指摘されている。

このような情報環境に関わる現象はエコーチェンバーやフィルターバブル現象として知られている (Nguyen et al. 2014; Cinelli et al. 2021)。エコー

チェンバーは似た意見をもつ人々の間で情報が繰り返し共有されることで、異なる立場の情報や反論に触れる機会が減少する状況を指す。フィルターバブルはアルゴリズムによって個人の関心に合う情報が優先的に提示されることで、利用者が自分の選好に合った情報環境の中に閉じ込められる状況を指す。これらの現象はそれ自体がただちに社会的問題を引き起こすわけではない。しかし、情報摂取の偏りが続くことで、意見の過激化や相互不信が強まる可能性がある。特に政治的対立が深刻化している社会では、オンライン空間における情報摂取の偏りが、現実社会における対立や不信を増幅する要因となりうる。そこで本研究では、オンライン情報空間における社会的分断の主要な課題の一つとして「意見分極化」に着目する。

意見分極化とは、各個人がもつ意見が単に多様である状態ではなく、異なる立場をもつ集団の間で意見の隔たりが大きくなり、中間的な意見や相互理解が失われていく過程を指す。オンライン空間では、同じ意見をもつ利用者同士の相互作用が強まりやすい一方で、異なる意見をもつ利用者に対しては反発や不信が生じやすい。その結果、利用者は自らの意見を補強する情報や他者との関係に継続的に接触し、集団間の意見差が維持または拡大される可能性がある (Chitra & Musco, 2020)。このような意見分極化の過程を理解するためには、個人の意見変化だけでなく、利用者間の関係性が友好的であるか敵対的であるかを含めて、意見形成のダイナミクスを捉える必要がある。

意見分極化を考えるうえでは、分極状態が形成された後の状態だけでなく、そこに至る意見形成の過程に注目することが重要である。現実のオンライン空間でも、人々の意見は一度に固定されるわけではなく、他者との相互作用を通じて徐々に変化していく。したがって、意見形成や分極化がどのように進行するのかを理解するだけでなく、その途中でどのような働きかけを行えば分極化を抑えられるのかを検討する必要がある。特に、敵対的関係を含む意見形成ダイナミクスでは、意見が分断された状態へ進みやすいことが知られている (Li et al., 2015)。このような状況に対しても、ネットワーク構造や各個人の意見状態を手がかりに小規模な介入を行うことで、分極化の進行を抑えられる可能性がある。

そこで本研究では、分極化へ向かう意見形成ダイナミクスを対象に、ネットワークへの小規模な介入によって分極化を抑制し、合意形成を促す方法を検討する (Chujyo et al., 2025)。具体的には、友好的関係と敵対的関係を同時に表現できる符号付きネットワーク上の投票者モデルを用いる。このモデルでは各個人の意見が周囲の人々との関係に影響を受けながら変化し、敵対的関係が存在することで分極状態が生じる場合がある。本研究では、このような意見形成過程に対して、符号構造と意見状態を考慮した小規模なネットワーク摂動（つながりの影響力に関する小規模な変更）を導入する。そして、実世界のデータで観察された符号付きネットワークに提案手法を適用し、敵対的関係が存在する場合でも分極化の抑制や合意形成が可能なのか、これらに必要な摂動量（変更の規模）はどの程度のものであるのかを評価する。これにより、分極化が生じる意見形成過程において、どの程度の小さな介入によって分断を抑制できるのかを数理的に明らかにする。

本研究が目指すのは、意見の違いそのものをなくすことではなく、違いが深刻な分断へ発展する過程をどのように抑えられるかを明らかにすることである。オンライン空間では、対立する人々の関係や情報の流れの一部を少し変えるだけでも、全体の意見の動きが変わる可能性がある。本研究では、このような「小さな働きかけ」が分断の拡大を防ぎ、より建設的な議論や合意形成につながる条件について数理モデルを用いて探る。

2. 意見形成モデル

意見形成モデルとは、人々が周囲の人々から影響を受けながら意見を変化させていく過程を数理的に表現するためのモデルである。現実社会では、人の意見は一人で独立に決まるわけではなく、家族、友人、同僚、ソーシャルメディア上のつながり、ニュースや推薦システムなど、さまざまな情報源や人間関係の影響を受けて変化する。意見形成モデルでは、このような複雑な社会的相互作用を単純化し、個人をネットワーク上のノード、人間関係や情報接触をエッジ（リンク）として表す。そして、各ノードが周囲のノードの意見に応じて自分の意見を更新するものとして、集団全体の意見が時間とともにどのように変化するのかを調べる。

このようなモデルは現実社会をそのまま再現するものではない。しかし、個人同士の単純な相互作用から合意形成、意見の共存、少数派意見の存続、分極化といった集団的な現象がどのように生じるのかを理解するうえで有用である。たとえば、各個人が周囲の人々に少しずつ同調するだけでも、ネットワーク構造によっては社会全体が一つの意見にまとまる場合もあれば、複数の意見が長く残る場合もある。このように、意見形成モデルは、個人レベルの相互作用と社会全体の意見分布を結びつけるための基本的な枠組みである。

2.1 投票者モデル

代表的な意見形成モデルの一つに「投票者モデル」がある（Holley & Liggett, 1975; Sood et al. 2008）。投票者モデルでは、各ノードが二つの意見のいずれかをもち、各時刻において隣接するノードの意見を参照し、その意見を採用する。これは「周囲の人の意見をまねる」という単純なルールによるモデルである。この単純さにもかかわらず、投票者モデルは局所的な模倣が社会全体の合意へとつながる過程を調べるための基本モデルとして広く用いられてきた。また、正則格子、スモールワールド・ネットワーク、スケールフリー・ネットワーク、多層ネットワークなど、さまざまなネットワーク構造上で解析されており、ネットワークの形が合意形成の速度や最終的な意見状態に大きく影響することが示されている。

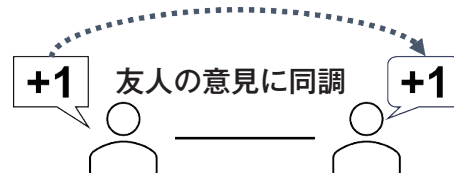
一方で、古典的な投票者モデルでは、人々が基本的に周囲の意見に同調することを仮定している。そのため意見の違いが一時的に存在していても、相互作用を繰り返すうちに、ネットワーク全体でどちらか一方の意見が広がりやすい。結果として、有限で連結なネットワークでは、長時間後には全員が同じ意

見を持つ合意状態に到達しやすく、二つの集団が異なる意見を安定して持ち続けるような分極状態は生じにくい。これは、古典的な投票者モデルが意見形成の基本的な仕組みを理解するうえで有用である一方、現実のオンライン空間で見られる持続的な対立や社会的分断を表すには十分ではないことを意味している。

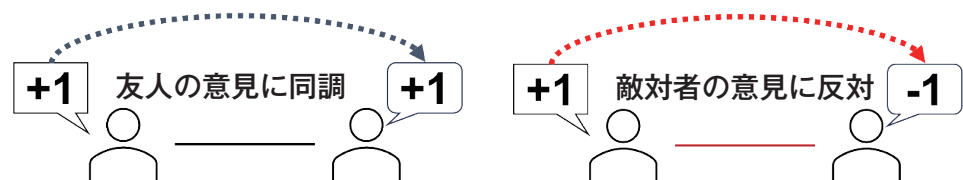
意見の分極化が進む要因には、情報推薦、選択的接触、感情的対立など、さまざまなものが考えられる。その中でも本研究では、人々の間に存在する友好的関係と敵対的関係に着目する。現実の意見形成では、人は必ずしも相手に同調するとは限らず、信頼している相手の意見には近づく一方で、不信感や敵対感をもつ相手に対しては、むしろ反対の意見を強めることもある。このような関係性を考慮することで、単なる同調だけでは説明しにくい分極状態を、意見形成モデルの中で表現できるようになる。

図表1 投票者モデルと符号付き投票者モデルの比較

投票者モデル



符号付き投票者モデル



出所) 筆者作成

2.2 友好的関係と敵対的関係を考慮した符号付き投票者モデル

友好的関係と敵対的関係を同時に扱うためには、関係の種類を区別できるネットワーク表現が必要である。本研究では、そのための枠組みとして符号付きネットワークを用いる (Li et al., 2015)。符号付きネットワークでは、各エッジに正または負の符号を与える。正のエッジは相手の意見に同調しやすい友好的な関係を表す。一方、負のエッジは相手の意見とは反対の方向に動きやすい敵対的な関係を表す。これにより、単に「誰と誰がつながっているか」だけでなく、「その関係が同調を生むのか、それとも反発を生むのか」を区別して扱うことができる。本研究で想定する符号付きネットワークと意見更新の基本的な考え方を図表1に示す。

本研究で用いる「符号付き投票者モデル」では、各ノードは二つの意見のいずれかを持つ。ここでは、意見を +1 または -1 の二値で表す。たとえば、ある論点に対する賛成と反対、あるいは二つの異なる立場を表していると考えることができる。各時刻において、ノードは周囲のノードの意見と、そのノード

との関係が友好的か敵対的かをもとに、自分の意見を維持するか反転させるかを決定する。

このモデルの基本的な考え方は、「友好的な相手とは同じ意見を持ちやすく、敵対的な相手とは反対の意見を持ちやすい」というものである。すなわち、友好的な相手が自分と同じ意見を持っている場合、その関係は現在の意見を支持する方向に働く。一方、友好的な相手が自分と異なる意見を持っている場合、その関係は自分の意見を変える方向に働く。敵対的な相手については逆であり、敵対的な相手が自分と反対の意見を持っている場合には現在の自分の意見が支持され、同じ意見を持っている場合には自分の意見を変える圧力が生じる。

以下では、この考え方を数式で表す。ネットワークは符号付き隣接行列 $A \in \{-1, 0, +1\}^{N \times N}$ によって表される。ここで $A_{ij}=+1$ はノード i とノード j の間に友好的な関係があることを表し、 $A_{ij}=-1$ は敵対的な関係があることを表す。また $A_{ij}=0$ は両者の間にエッジが存在しないことを表す。各ノード i は時刻 t において意見状態 $x_i(t) \in \{-1, +1\}$ をもつ。各時刻において、すべてのノードは周囲のノードから受ける影響をもとに、自分の意見を反転させるかどうかを確率 $p_i(t)$ で決定する。ここで意見反転確率 $p_i(t)$ は以下のように表される。

$$p_i(t) = \frac{1}{2} \left[1 - \frac{x_i(t)}{\sum_{h=1}^N |A_{ih}|} \sum_{j=1}^N A_{ij} x_j(t) \right]$$

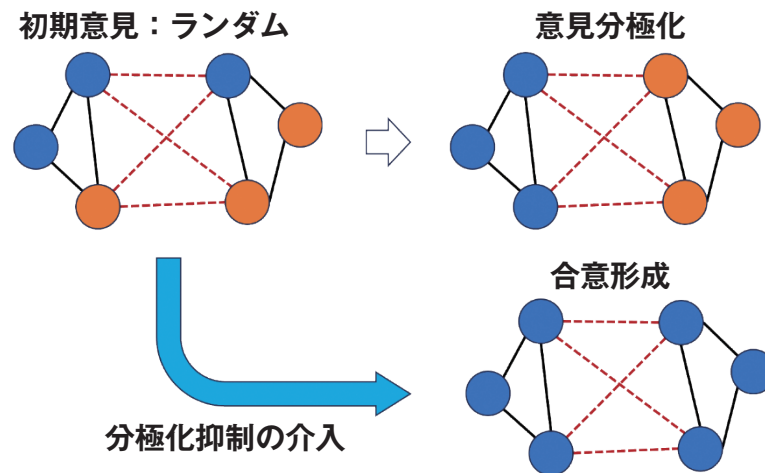
ここで $\sum_{j=1}^N A_{ij} x_j(t)$ はノード i が周囲から受ける全体的な影響を表している。友好的な隣人の意見はそのまま影響し、敵対的な隣人の意見は符号が反転して影響する。つまり、友好的な隣人と同じ意見を持ち、敵対的な隣人とは反対の意見を持っているほど、意見を変える確率は小さくなる。一方で、それらと整合しないほど、意見を変える確率は大きくなる。また分母の $\sum_{h=1}^N |A_{ih}|$ は、ノード i が持つ隣人の数を表しており、この項で正規化することで意見反転確率が適切な範囲に収まるようにしている。本研究では、すべてのノードが同時に意見を更新する同期更新を考える。

このモデルでは、ネットワーク構造によって最終的な意見状態が変化する。特に、同じ集団内に友好的関係が多く、異なる集団間に敵対的関係が多い場合、各集団の内部では意見がそろいやすく、集団間では反対の意見が維持されやすい。その結果、ネットワーク全体が一つの意見にまとまるのではなく、二つの集団が異なる意見を持つ分極状態が生じやすくなる。このような状態は、符号付きネットワークにおける符号バランスの考え方と関係している。符号バランスとは「友人の友人は友人であり、敵の敵は友人である」といった関係がネットワーク全体で矛盾なく成り立つ状態である。このような構造では、各集団が互いに反対の意見を持つ分極状態が、ネットワーク構造と合致した状態となる。

本研究では、このように分極化へ向かいやすい符号付き意見形成モデルを対象とし、その過程に小規模なネットワーク介入を加えることで、分極化の進行

を抑制できるかを調べる。ここでの介入とは、各個人の意見そのものを直接変えるのではなく、ネットワーク上の一部の関係性に働きかけることを意味する。図表2に、分極化へ向かう意見状態の推移と、それに対する介入のイメージを模式的に示す。

図表2 符号付き意見形成モデルにおける分極化と合意形成の概念図



出所) 筆者作成

オレンジと青は、それぞれ異なる意見を示し、黒線と赤点線は友好的関係と敵対的関係を表す。符号付き意見形成モデルでは、初期意見がランダムであっても、ネットワーク構造によって意見が分極化する場合がある。本研究では、このような分極化を抑制するためにネットワーク介入を行い、合意形成へと近づけることを目指す。

3. 意見形成過程への介入方策

意見形成モデルは、分極化が生じる仕組みを説明するためだけでなく、分極化をどのように抑制できるのかを検討するためにも用いることができる。近年では、意見形成の過程にどのような働きかけを行えば、合意形成を促したり分極化を緩和したりできるのかという観点から、意見形成モデルへの介入方策が注目されている。意見形成の過程がネットワーク上の相互作用として表されるならば、個人の意見や人々の関係性に働きかけることで、集団全体の意見の動きを変えられる可能性がある。

意見形成への介入にはいくつかの考え方があり、たとえば、特定の個人の初期意見を変える、影響力の大きい人物を選んで情報を発信させる、人々の説得されやすさを変える、推薦システムを通じて接触する情報を調整する、といった方法が考えられる。また、ネットワーク構造そのものに働きかける方法もある。これは関係の有無や強さ、異なる集団間の接触のあり方を調整することで、意見の広がり方や分極化の進み方を変えようとするものである。

本研究では、これらの介入の中でも、ネットワーク構造に対する小規模な介入に着目する。各個人の意見を直接変えるのではなく、意見形成を支えている関係性に小さな変化を加えることで、分極化へ向かう意見ダイナミクスを変化させることを目指す。具体的には、投票者モデルに対して提案された敵対的摂動による介入手法（Chiyomaru & Takemoto, 2022; Chiyomaru & Takemoto, 2023; Mizutaka, 2023）をもとに、友好的関係と敵対的関係を含む符号付き投票者モデルへ拡張する。次節ではこの介入手法を具体的に定式化する。

3.1 提案手法：敵対的摂動に基づく分極化抑制介入手法

ここでは敵対的摂動に基づく介入手法を、符号付き投票者モデルに適用する方法を説明する。この手法は、機械学習における敵対的攻撃の考え方を意見形成モデルに応用したものである。敵対的攻撃とは、機械学習モデルの入力に人間にはほとんど分からない程度の小さな変化を加えることで、モデルの出力を意図的に変化させる手法である（Goodfellow et al., 2015）。意図的なコントロールを試みるものであるため、敵対的攻撃と呼称されている（前出の意見の敵対とは異なる文脈で用いられる用語である）。これを意見形成モデルに置き換えると、ネットワーク上の関係性に小さな変化を加えることで、その後の意見ダイナミクスを望ましい方向へ誘導する操作として解釈できる（これを敵対的摂動と呼ぶ）。つまり、各個人の意見を直接変更しなくても、どの関係にどの向きの摂動を加えるべきかを適切に選ぶことで、分極化へ向かう流れを弱められる可能性がある。

既存の意見形成モデルに対する敵対的摂動の枠組みでは、主に同調的な関係を前提としていた（Chiyomaru & Takemoto, 2022; Ninomiya et al., 2025）。これに対して、本研究では友好的関係と敵対的関係を同時に含む符号付きネットワークを対象とする。敵対的関係が存在する場合、ある関係を強めることが必ずしも合意形成を促すとは限らない。友好的な相手からの影響を強めるべき場合もあれば、敵対的な相手からの影響を弱めるべき場合、あるいは敵対的な関係を利用することで目標意見への誘導が進む場合もある。そのため、符号付きネットワークでは、関係が友好的なのか敵対的なのかを考慮して摂動を設計する必要がある。

以下では、目標とする合意状態を定め、その状態に意見ダイナミクスを近づけるための摂動を数理的に定義する。そのために、現在の意見更新が目標状態からどの程度離れているかを表すエネルギー関数を導入する。このエネルギーを小さくする方向にネットワーク上の関係性をわずかに変化させることで、分極化へ向かう意見形成過程を合意状態へ誘導する。

ここでは各ノード i の次時刻の意見 $x_i(t+1)$ を目標状態 x^* に近づけることを考える。合意形成を目標とするため、すべてのノードに対して同じ目標意見 x^* を設定する。たとえば、 $x^*=+1$ は全ノードが最終的に $+1$ の意見へ近づくことを目指すことを意味する。このとき、次時刻の意見 $x_i(t+1)$ が目標状態 x^* にどの程度近いかを評価するために、次のエネルギー関数を定義する。

$$E(t) = -\frac{1}{N} \sum_{i=1}^N x^* x_i(t+1)$$

このエネルギーは、現在の意見ダイナミクスが目標状態にどの程度近づいているかを表す指標である。各ノードの意見が目標状態と一致していると $x^* x_i(t+1)$ は正になり、エネルギー $E(t)$ は小さくなる。したがって、エネルギーを小さくするようにネットワークを変化させれば、意見状態を目標とする合意状態へ近づけることができる。

次に、どの関係性にどの向きの変化を加えれば、このエネルギーを小さくできるかを考える。そのために用いるのが、エネルギー $E(t)$ の A_{ij} に関する勾配である。ここでの勾配 $\partial E(t)/\partial A_{ij}$ はノード i とノード j の関係 A_{ij} をわずかに変化させたとき、エネルギーがどの方向にどれだけ変化するかを表す。したがって、勾配の符号を見ることで、各関係性をどちらの向きに変化させればエネルギーが小さくなり、目標状態に近づきやすいかを判断できる。

この考え方は勾配降下法に対応している。勾配降下法では、ある量を小さくしたいとき、その量が最も増える方向とは反対向きに変数を動かす。本研究では、エネルギー $E(t)$ を小さくするために、符号付き隣接行列の各要素 A_{ij} を勾配と反対向きにわずかに変化させる。摂動後の隣接行列の要素は次のように表される。

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

ここで、 $\epsilon > 0$ は摂動の大きさを表すパラメータである。 ϵ が大きいほど、各エッジに対する介入は強くなる。本研究では、ネットワーク全体の繋がり方を作り替えるのではなく、すでに存在するエッジに対してのみ摂動を加える。上式は、エネルギーを小さくするための基本的な考え方を表している。

一方で、意見更新は確率的に行われるため、実際の $x_i(t+1)$ は毎回の試行によって変化する。そのため、エネルギーの勾配をそのまま扱うことは難しい。そこで本研究では、時刻 t から $t+1$ への更新を平均的な変化として近似する。これは、意見更新の時間発展を平均場的に扱うことに対応する。具体的には、ノード i の次時刻の意見を、意見を反転する場合と維持する場合の期待値として表す。これにより次のように近似できる。

$$x_i(t+1) = -x_i(t)p_i(t) + x_i(t)(1-p_i(t))$$

これによりエネルギー $E(t)$ が各エッジの変化に対してどのように増減するかを計算できる。

本研究で必要となるのは、勾配の大きさそのものではなく、各関係性をどちらの向きに変化させれば目標状態に近づきやすいかである。そこで、機械学習における Fast Gradient Sign Method と同様に、勾配の符号のみを用いる簡

略化を考える (Goodfellow et al., 2015)。これにより、計算を単純化しながら、エネルギーを減少させる方向への摂動を決定できる。この考え方に基づくと、摂動後の隣接行列の要素は次のように表される。

$$A_{ij}^* = A_{ij} - \epsilon \times \text{sign}\left(\frac{\partial E(t)}{\partial A_{ij}}\right)$$

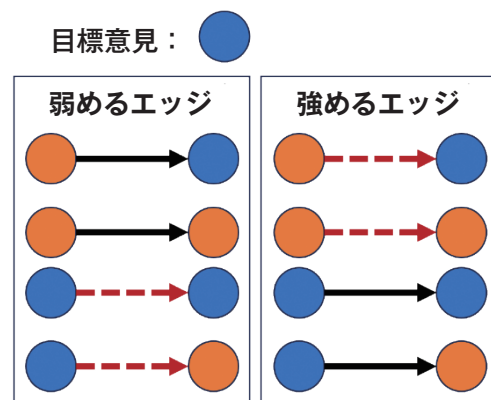
ここで、 $\text{sign}(x)$ は x の符号を返す関数である。この式は、勾配が正であれば A_{ij} を小さくし、勾配が負であれば A_{ij} を大きくすることを意味する。つまり、エネルギーが小さくなる方向にだけ注目して、各関係性に小さな摂動を加える。さらに、符号付き投票者モデルに対してこの式を整理すると、摂動は最終的に次の単純な形で表される。

$$A_{ij}^* = A_{ij} + \epsilon x_j^* x_j(t)$$

この式は、各時刻において、既存のエッジ A_{ij} に小さな摂動 $\epsilon x_j^* x_j(t)$ を加えることを表している。隣人が目標意見と同じ意見を持つ場合には関係を正の方向に変化させ、反対の意見を持つ場合には負の方向に変化させる。したがって、各時刻の意見状態に応じて、どの関係をどちらの向きに調整すべきかを単純な規則で決めることができる。このような摂動を各時刻で加えながら意見更新を行うことで、分極化へ向かう意見形成過程を目標とする合意状態へ近づけることを目指す。

図表3は、この式が表す介入の考え方を模式的に示している。目標とする意見を青とした場合、青の意見を持つ相手との関係は、目標意見を後押しする方向に調整される。一方、オレンジの意見を持つ相手との関係は、目標意見から離れる影響を抑える方向に調整される。つまり、この介入では、どの相手からの影響を強め、どの相手からの影響を弱めるかを調整することで、全体の意見状態を合意に近づける。

図表3 提案手法のネットワーク介入の概念図



出所) 筆者作成

オレンジと青は異なる意見を示しており、ここでは青を目標意見としている。また黒線は友好的関係を、赤の点線は敵対的関係を表す。青の意見を持つ相手との関係は目標意見を後押しする方向に調整され、オレンジの意見を持つ相手との関係は目標意見から離れる影響を抑える方向に調整される。このように、各時刻の意見状態に応じて関係性を調整することで、合意形成に近づけることを目指す。

4. 実際のネットワークデータにおける分極化抑制の評価

本節では、提案した敵対的摂動に基づく介入手法が、実世界の符号付きネットワークにおいて分極化の進行を抑制し、合意状態への接近を促すかを検証する。とくに意見状態とネットワークの符号構造を考慮して摂動を加える提案手法と、同じ強度のランダム摂動を比較することで、構造化された介入がどの程度有効であるかを評価する。

4.1 評価の目的と指標

ここでの評価の目的は、分極化へ向かいやすい符号付き投票者モデルに対して、小規模なネットワーク摂動を加えることで、意見状態をどの程度合意状態へ近づけられるかを調べることである。目標状態をすべてのノードが意見 $x^* = +1$ を持つ状態とし、介入後の意見ダイナミクスがこの目標状態にどの程度近づいたかを評価する。

合意状態への近さは、最終時刻において目標意見を持つノード割合である合意割合 $\langle \rho \rangle$ によって測定する。たとえば、 $\langle \rho \rangle = 1$ であれば、すべてのノードが目標意見 x^* を持つ完全な合意状態を意味する。一方、 $\langle \rho \rangle = 0.5$ であれば、目標意見を持つノードとそうでないノードがほぼ半々に分かれており、本モデルでは分極状態に近い状態として解釈できる。本研究では終端時刻を $t_{\max} = 1,000$ とし、その時点での $\langle \rho \rangle$ を用いて介入効果を評価する。

介入効果の比較手法として、上述の提案手法と同じ強度を持つランダム摂動を用いる。提案手法では、各時刻の意見状態と目標意見に基づいて、既存のエッジに対して方向を定めた摂動を加える。一方、ランダム摂動では、各エッジに対して正または負の摂動をランダムに加える。両者を比較することで、単にネットワークにノイズを加えるだけで分極化が緩和されるのか、それとも意見状態を考慮した構造化された摂動が必要なのが検証可能となる。

4.2 使用するデータセット

提案手法の有効性を検証するために、本研究では Stanford Network Analysis Project (SNAP) で公開されている七つの符号付きネットワークの実データを使用する (Leskovec & Krevl, 2014)。この七つのネットワークデータは、いずれも現実のオンラインサービス上で観測された、人・コミュニティ間のつながりをグラフ化したデータである。これらのデータセットは肯定的な関係だけではなく、否定的な関係も明示的に記録されている点に特徴があ

る。単にユーザ間に関係があるかどうかだけでなく、その関係が肯定的なものか否定的なものか、あるいは信頼を表すものか不信を表すものかを区別することができる。本研究では感情、友人・敵対関係、信頼・不信頼、評判といった異なる意味をもつ符号付き相互作用を含む複数のプラットフォームを対象とした。また、ネットワークの規模、疎密性、負の結合の割合が異なるデータセットを用いることで、提案手法が多様な実世界ネットワークに対して有効であるかを評価する。

使用したデータセットの概要を図表4に示す。Reddit Hyperlink ネットワークでは、subreddit のノード間のリンク関係を扱う（Reddit は英語圏を中心とした大規模な掲示板・SNS であり、テーマごとに subreddit がある）。ある投稿が別の投稿へリンクしている場合に subreddit のノード間にエッジが張られ、そのリンク元投稿の感情に基づいて符号が与えられる。本研究では、中立または肯定的な感情を +1、否定的な感情を -1 とした。Slashdot Zoo の三つのスナップショットでは、ユーザ間の関係が友人関係 +1・敵対関係 -1 として記録されている（Slashdot はテクノロジー系の掲示板）。Bitcoin Alpha および Bitcoin OTC は、Bitcoin トレーダー間の評判スコアを記録したネットワークであり、正の評価を信頼 +1、正でない評価を不信頼 -1 とした。Epinions は消費財に関するレビューア間の信頼関係を表すネットワークであり、信頼を +1、不信頼を -1 として扱った。

すべてのデータセットについて、エッジ重みは用いず、符号のみを扱った。また、双方向にエッジが存在する場合には、多数派の符号をもつ単一の無向エッジへ集約し、孤立ノードは除去した。これにより、各データセットを無向の符号付き隣接行列として使用する。

図表4 使用する実ネットワークデータの統計量

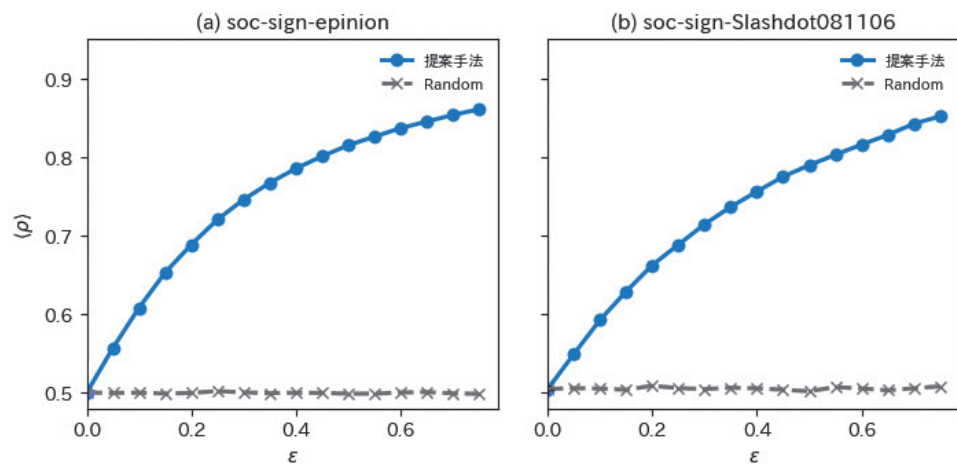
データ名	ノード数	エッジ数	正エッジ数	負エッジ数
soc-RedditHyperlinks	35,776	286,561	265,491	21,070
soc-sign-Slashdot081106	77,350	516,575	396,378	120,197
soc-sign-Slashdot090216	81,867	545,671	422,349	123,322
soc-sign-Slashdot090221	82,140	549,202	425,072	124,130
soc-sign-bitcoinalpha	3,783	24,186	22,650	1,536
soc-sign-bitcoinotc	5,881	35,592	32,029	3,563
soc-sign-epinion	131,828	841,372	717,667	123,705

出所) Chuiyo et al. (2025) の Table 3 を日本語名に変更したもの

4.3 提案手法の分極化抑制の評価

図表5は、実世界符号付きネットワークにおいて摂動強度 ϵ を変化させたときの合意割合 $\langle\rho\rangle$ の変化を示している。提案手法である敵対的摂動では、 ϵ の増加に伴って $\langle\rho\rangle$ が一貫して上昇する。一方、同じ強度のランダム摂動では、 ϵ を大きくしても $\langle\rho\rangle$ はほとんど改善せず、多くの場合で分極状態に近い値にとどまる。これは、単にネットワークにランダムなノイズを加えるだけでは分極状態を解消しにくく、ネットワーク構造と意見状態に基づいて摂動の向きを定めることが重要であることを示している。

図表5 実世界の符号付きネットワークにおける性能



出所) Chuijo et al. (2025) Figure 7を日本語ラベルに変更したもの。

注) (a) Epinionおよび (b) Slashdot081106における意見合意割合 $\langle\rho\rangle$ と摂動強度 ϵ の関係を示している。青丸が提案した敵対的摂動を示し、灰色の十字はランダムノイズを示す。

図表6は、代表的な摂動強度における合意割合を示している。摂動を加えない場合 ($\epsilon=0$) では、多くのネットワークで $\langle\rho\rangle$ は0.5付近にとどまる。これは、目標意見をもつノードとそうでないノードがほぼ半々に分かれており、分極状態に近いことを意味する。これに対して、提案手法による摂動を加えると、すべてのデータセットで $\langle\rho\rangle$ は上昇する。

ネットワークごとに見ると、Bitcoin AlphaとBitcoin OTCでは、比較的小さな摂動でも合意割合が大きく上昇しており、提案手法の効果がとくに顕著である。たとえば、 $\epsilon=0.1$ の時点で、Bitcoin Alphaでは $\langle\rho\rangle=0.825$ 、Bitcoin OTCでは $\langle\rho\rangle=0.812$ に到達している。これは、これらのネットワークでは、わずかな構造化摂動によって意見ダイナミクスを目標状態へ誘導しやすいことを示している。Reddit Hyperlinksでも、 $\epsilon=0.1$ で $\langle\rho\rangle=0.711$ 、 $\epsilon=0.2$ で $\langle\rho\rangle=0.814$ となり、比較的小さな介入で高い合意割合に到達する。

一方で、Slashdotの三つのスナップショットとEpinionsでは、合意割合の上昇はより緩やかである。たとえば、Epinionsでは $\epsilon=0.1$ で $\langle\rho\rangle=0.608$ 、 $\epsilon=0.2$ で $\langle\rho\rangle=0.688$ にとどまる。また、Slashdotの各スナップショットでも、 $\epsilon=0.2$ における $\langle\rho\rangle$ はおおむね0.66前後である。これらのネットワークでは、敵対的関係や符号構造がより複雑に配置されており、分極状態を大きく

動かすにはより強い摂動が必要となっている可能性がある。

それでも、摂動強度を十分に大きくすると、すべてのネットワークで合意割合は大きく上昇する。 $\epsilon=0.7$ では、最も上昇が緩やかな Slashdot や Epinions でも $\langle\rho\rangle$ は 0.84 以上に到達している。このことは、実世界の複雑な符号付きネットワークであっても、意見状態と符号構造を考慮した摂動を加えることで、分極化へ向かうダイナミクスを合意状態へ大きく近づけられることを示している。

以上の結果は、ネットワークモデルで得られた傾向とも整合している (Chujiyo et al., 2025)。すなわち、ランダムなエッジ摂動は分極状態をほとんど解消できない一方で、符号と意見状態を考慮した摂動は、実世界の不均質な社会ネットワークにおいても合意割合を着実に高める。また、その効果の大きさはネットワークごとに異なり、負のエッジの割合や符号構造の整合性が、必要な摂動強度を左右する可能性がある。

図表6 実データにおいて摂動強度 ϵ を変えた結果

データ名	$\epsilon=0$	$\epsilon=0.1$	$\epsilon=0.2$	$\epsilon=0.7$
soc-RedditHyperlinks	0.500	0.711	0.814	0.938
soc-sign-Slashdot081106	0.505	0.593	0.662	0.842
soc-sign-Slashdot090216	0.504	0.587	0.657	0.841
soc-sign-Slashdot090221	0.500	0.590	0.657	0.841
soc-sign-bitcoinalpha	0.590	0.825	0.904	0.981
soc-sign-bitcoinotc	0.551	0.812	0.876	0.967
soc-sign-epinion	0.500	0.608	0.688	0.853

出所) Chujiyo et al. (2025) の Table 1 を日本語名に変更したもの。

5. 議論と今後の課題

本研究では、友好的関係と敵対的関係を含む符号付き投票者モデルに対して、ネットワーク摂動を加えることで、分極化を抑制する手法を検討した。実世界の符号付きネットワークデータを用いた数値実験の結果、提案手法はランダムな摂動と比べて、目標意見を持つノードの割合を一貫して高めることが明らかになった。この結果は、分極化した意見状態を変えるためには、意見状態と関係性の符号を考慮した介入方策が重要であることを示している。

本結果は、敵対的関係が多数存在するネットワークでは合意形成がより難しくなることも示唆している。従来の投票者モデルのようにすべての関係が同調的に働く場合には、比較的小さな摂動でも意見状態を目標方向へ誘導しやすいことが知られている (Chiyomaru & Takemoto, 2022)。一方、本研究で扱った符号付きネットワークでは、敵対的関係が相手の意見に反発する方

向に働くため、分極状態が維持されやすい。そのため、敵対的関係が多いネットワークや複雑な符号構造を持つネットワークでは、合意状態へ近づけるために、より大きな摂動が必要になる。この点は、オンライン空間における敵対的関係や相互不信が、分極化の緩和を難しくする要因となりうることを示している。

本研究で扱う介入は、ユーザ間の関係そのものを直接変更することを意味するものではない。実際のオンラインサービスにおいて、人間関係や信頼関係を外部から直接書き換えることは、現実的でないだけでなく、望ましい介入ともいえない。むしろ、本研究における摂動は、ランキング、推薦、検索結果の提示、評判スコアの集約などを通じて、情報や他者から受ける影響の重みをわずかに調整する操作として解釈できる。たとえば、対立を強める情報接触の影響を弱めることや、異なる立場のユーザ同士が過度に敵対的にならないかたちで接触できるようにすることが考えられる。この意味で、本研究の結果は分極化を抑制するためのアルゴリズム的介入を設計するうえで、基礎的な数理的知見を与えるものである。

一方で、このような介入を実社会に応用する際には、慎重な議論が必要である。意見形成過程への介入は、たとえ分極化の緩和を目的とするものであっても、利用者にとっては情報環境が見えにくいかたちで調整されることを意味する。透明性を欠いた介入は、プラットフォームに対する不信を高めたり、意見誘導や検閲と受け取られたりする可能性がある。また、同じ仕組みが不適切に用いられれば、分極化を緩和するのではなく、特定の方向へ世論を誘導したり、逆に分断を強めたりする目的で利用される危険性もある。したがって、実際の応用を考える際には、介入の目的、対象、方法を明確にするとともに、透明性、利用者の選択可能性、公平性、説明可能性、悪用防止の仕組みを十分に検討する必要がある。

最後に、本研究の限界についていくつか述べる。第一に、本研究では単純化された二値意見モデルを用いた。しかし、現実の意見は連続的または多次元的である場合が多く、個人の意見には確信度、関心の強さ、文脈依存性なども伴う。したがって、今後は連続値意見モデルや多次元意見空間をもつモデルへ拡張し、提案手法がより複雑な意見形成過程においても有効であるかを検討する必要がある。

第二に、本研究ではネットワーク構造を固定したうえで、既存の関係に摂動を加える設定を考えた。しかし、実際のオンライン空間では、ユーザ間の関係や情報接触は時間とともに変化する。そのため、時間発展するネットワークや多層ネットワークにおいて、どのような介入が分極化の抑制に有効であるかを検討することが今後の課題である。特に、社会的関係、情報拡散経路、推薦システムによる接触機会が相互に作用する状況を扱うことは、現実のオンライン情報空間を理解するうえで重要である。

第三に、本研究ではネットワーク全体の構造や各ノードの意見状態が観測可能であることを前提としている。しかし、実際のプラットフォームでは、全ユーザの関係性や意見状態を完全に観測できるとは限らない。また、意見状態の推定には誤差が含まれ、観測可能な関係もプラットフォーム上の一部の相互

作用に限られる場合がある。したがって、部分的な観測情報や不確実な意見推定に基づいても有効に機能する介入手法を開発することが重要である。

本研究は数理モデルと実データに基づく数値実験にとどまっており、実際の人間集団に対して介入を行ったものではない。そのため、提案手法が現実のオンライン空間でどのように機能するかを評価するには、慎重に設計された実験や、倫理的配慮を伴う検証が必要である。また、分極化の抑制を目指す際には、合意形成を促すことと多様な意見を維持することのバランスも重要である。本研究の目的は、すべての意見を一つにそろえることではない。何が望ましい状態かの社会的合意は、上述のような配慮がなされたとしても容易ではない。しかし、社会分断が過度に強まり、まっとうな議論さえ成立しなくなる状況が望ましい状態ではないことには広く合意が得られるであろう。本研究の目的は、こうした敵対的な分断が過度に強まる過程をどのように抑えられるかを理解することである。

6. 結論

本研究では、敵対的関係を含む符号付きネットワークにおいて、意見状態と関係性の符号を考慮した小規模な摂動介入により、分極化の進行を抑制できる可能性を示した。提案手法は、分極化を単に観察するだけでなく、その緩和可能性を数理的に評価するための枠組みを提供するものである。今後は、より現実的な意見モデル、時間変化するネットワーク、部分観測下での介入設計へと拡張が求められる。さらに、倫理的に配慮された実証的検証と組み合わせることで、オンライン情報空間における社会的分断の緩和に向けた透明で慎重な介入設計のあり方を明らかにしていく必要がある。

参考文献

- Chitra, U., & Musco, C. (2020). Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th international conference on web search and data mining*, 115-123.
- Chiyomaru, K., & Takemoto, K. (2022). Adversarial attacks on voter model dynamics in complex networks. *Physical Review E*, 106(1), 014301.
- Chiyomaru, K., & Takemoto, K. (2023). Mitigation of adversarial attacks on voter model dynamics by network heterogeneity. *Journal of Physics: Complexity*, 4(2), 025009.
- Chuijo, M., Liu, S., & Toriumi, F. (2025). Steering polarization toward consensus in signed majority-vote opinion models. *Scientific Reports*, 15(1), 39882.
- Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. In *Proceedings of the national academy of sciences*, 118(9), e2023301118.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. In Y. Bengio & Y. LeCun (Eds), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May*

- 7-9, 2015, *Conference Track Proceedings*.
- Holley, R. A., & Liggett, T. M. (1975). Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, 643-663.
- Leskovec, J., & Krevl, A. (2014, June). SNAP Datasets: Stanford large network dataset collection. <https://snap.stanford.edu/data/>
- Li, Y., Chen, W., Wang, Y., & Zhang, Z. L. (2015). Voter model on signed social networks. *Internet Mathematics*, 11(2), 93-133.
- Mizutaka, S. (2023). Crossover phenomenon in adversarial attacks on voter model. *Journal of Physics: Complexity*, 4(3), 035009.
- Nguyen, T. T., Hui, P. M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, 677-689.
- Ninomiya, M., Ichinose, G., Chiyomaru, K., & Takemoto, K. (2025). Mitigating opinion polarization in social networks using adversarial attacks. *Scientific Reports*, 15(1), 9033.
- Sood, V., Antal, T., & Redner, S. (2008). Voter models on heterogeneous networks. *Physical Review E-Statistical, Nonlinear, and Soft Matter Physics*, 77(4), 041121.