

SNS 上に氾濫する不適切コンテンツへの注目度抑制手法の確立

代表研究者

平 倉 直 樹

富山県立大学情報工学部 助教

1 はじめに

ソーシャルメディアは情報流通の主要な基盤となり、アテンションエコノミーと呼ばれる原理に支配されている[1]–[3]。アテンションエコノミーとは、利用者の注目そのものが経済的価値をもつ状況を指す。ソーシャルメディアプラットフォームの収益の大半は広告に由来するため、利用者の利用時間を最大化することが事業上不可欠である。この目的のため、プラットフォームは多くの注目を集める発信者を優遇してコンテンツ投稿を促し[4]、その結果として豊富なコンテンツが提供される。一方で、近年はその負の側面が顕在化している。利益さえ得られればどのようなコンテンツでも構わないという発想のもと、注目を集めるために不適切なコンテンツを投稿する発信者が増加している。具体的には、真実よりもインパクトを優先する偽情報[5]、[6]や、意図的に炎上を誘発するコンテンツ[7]、ヘイトスピーチ[8]などである。こうした不適切コンテンツの拡散は、プラットフォームの価値と利用者体験を損なうだけでなく、世論の分極化[9]や社会的対立の激化を招きうる。

不適切コンテンツへの対策は、主に影響最小化問題の枠組みで研究されてきた[10]–[12]が、これらの多くはアカウントの削除や制限といったモデレーションを前提とする。モデレーションには公平性・透明性の欠如や、特定の意見・コミュニティが過度に抑制される懸念が指摘されており[13]、自由な表現や多様性を損なうおそれがある。また、影響最小化・影響制限[14]のいずれの手法も、一般利用者を含む大規模なネットワークを考慮する必要があるという共通の課題を抱える。

本研究では、これらの課題を回避する手法として、発信者間の比較的小規模な競争ネットワークに着目した模倣戦略を構築した。模倣戦略は、ある発信者が別の発信者のコンテンツを模倣して発信者間の競争を強めることで、注目を分散・抑制する介入手法である。ただし、単一のターゲットを対象とする模倣戦略には、(i) 発信者が不適切か否かを区別しない、(ii) ネットワーク全体の注目度を下げると、不適切でない一般発信者の注目度まで低下させてしまうという課題があった。そこで、これらの課題を補い、複数の不適切発信者への注目を選択的に抑制しつつ一般発信者への影響を最小化する最適模倣戦略を構築した。

2 アテンションダイナミクスモデルに基づく模倣戦略

2-1 アテンションダイナミクスモデル

情報発信者への注目度の時間変化を記述するため、生物種の個体数変動を表す Lotka-Volterra 競争方程式に視聴者の飽きの効果を組み込んだモデル[15]を用いた。各発信者を生物種に、発信者が獲得する注目度を個体数に対応づける。発信者 i ($i = 1, \dots, N$) の時刻 t における注目度を $a_i(t)$ とすると、モデルは

$$\frac{da_i(t)}{dt} = r_i a_i(t) \left(1 - \frac{(b_i(t) + \sum_j w_{ij} a_j(t))}{K_i} \right)$$

で表される。ここで r_i は内的な成長率、 w_{ij} は他の発信者 j との競争係数、 K_i は環境収容力である。 K_i は、ある話題やジャンルにおいて発信者が獲得しうる注目度の上限を表す。 $b_i(t)$ は視聴者の飽きを表す項であり、発信者が注目を集めるほど視聴者が飽きていくことを反映して、過去の注目度の畳み込み $b_i(t) = \int_0^t a_i(s) \exp(-\zeta_i(t-s)) ds$ として定義される。 ζ_i は飽きの減衰率で、値が小さいほど視聴者が飽きやすいことを意味する。競争係数 w_{ij} は発信者間のニッチ（対象とする視聴者層）の重なりを程度を表し[16]、類似したコンテンツを発信する発信者ほど同じ視聴者を奪い合うため競争係数は大きくなる。Leibniz の公式[17]により飽き項の時間微分は $db_i/dt = a_i(t) - \zeta_i b_i(t)$ と書ける。以上より、 N 人の発信者の注目度と飽きに関する $2N$ 次元の微分方程式系を考える。

2-2 注目度抑制のための模倣戦略

注目度を低下させる要因を明らかにするため、ネットワークの隣接行列のスペクトルに基づく次元縮約法

[18]を適用した．全ノードのパラメータが等しい ($r_i = r$, $K_i = K$, $\zeta_i = \zeta$) と仮定した簡略モデルに対して縮約を行うと，ネットワーク全体の注目度 A と飽き B に関する 2 次元の縮約系

$$\frac{dA}{dt} = rA \left(1 - \frac{B}{K}\right) - \lambda \frac{r\mu A^2}{K}, \quad \frac{dB}{dt} = A - \zeta B$$

が得られる．ここで λ は隣接行列の最大固有値， μ は $\mu = \frac{v^T D v}{\lambda v^T v}$ (D は次数行列) で与えられる．この縮約系の固定点では関係式

$$A = \frac{K\zeta}{1 + \zeta\mu\lambda}$$

が成り立つ．この式は，隣接行列の最大固有値 λ を増大させると全体の注目度 A が低下することを示している．一般に最大固有値はネットワークのエッジ密度やエッジ重みが増すと大きくなる．エッジ重みは発信者間の競争の強さ，すなわちコンテンツの類似度によって定まる．各発信者を特徴ベクトル c_i で表す[19]と，発信者間の類似度はコサイン類似度 $S(c_i, c_j)$ で定義され，隣接行列は $W_{ij} = \max(S(c_i, c_j), 0)$

($i \neq j$, 対角成分は 0) と定める．負の類似度は無関係なコンテンツとみなして 0 とする．

上記の関係から，最大固有値を増大させればネットワーク全体の注目度を抑制できる．そこで，ある発信者（模倣者，添字 i で表す）が別の発信者（ターゲット，添字 τ で表す）のコンテンツを模倣することで両者の競争を強め，エッジ重みと最大固有値を増大させる介入を「模倣戦略」として提案した．模倣により模倣者の特徴ベクトル $\hat{c}_i$ は， c_τ と同一のベクトルへ変化する．これによりエッジ重みが増大し，最大固有値が増加し，結果としてネットワーク全体の注目度が低下する．

模倣戦略は模倣者とターゲットの間の競争を強める一方，模倣者和其他のノードとの競争関係も変化させるため，必ずしも成功するとは限らない．そこで，最大固有値の変化 $\hat{\lambda} - \lambda > 0$ を成功の条件とし，行列摂動論[20]を用いてこの変化を評価した．摂動行列を P とすると，変化後の最大固有値は $\hat{\lambda} \approx \lambda + \epsilon v^T P v$ と近似でき，右辺第二項が正となると戦略は有効である．解析の結果，固有ベクトル中心性が介入の成否を左右する鍵であることが分かった．固有ベクトル中心性[21]は，エッジの本数を重視する次数中心性とは異なり，重要なノードと接続しているノードほど重要とみなす指標である．本研究の競争ネットワークでは，競争の激しいノードと競争しているノードほど重要とみなされることを意味する．導出された条件から，(i) 模倣者和其他ノードとの競争を最大化する模倣が望ましいこと，(ii) ターゲットの固有ベクトル中心性が高いほど介入効果が大きいことが示された．

2-3 数値実験

模倣戦略を検証するため，特性の異なる 3 種類のネットワーク（エッジ密度の低い疎なネットワーク，密なネットワーク，一部に密な領域をもつ不均質なネットワーク）を用いて数値実験を行った．各ネットワークについて模倣者とターゲットのすべての組合せに模倣戦略を適用し，最大固有値の変化を評価した．実験の結果を図 1 に示す．いずれのネットワークでも，固有ベクトル中心性の高いノードをターゲットに選ぶことが有効であった．一方，中心性の高いノードを模倣者とし中心性の低いノードをターゲットに選ぶと，最大固有値はむしろ低下した．これは既存の競争を放棄することによる．密なネットワークでは元々の競争が激しいため，模倣により放棄される競争の影響が，新たに生じる競争の効果を上回る場合が多く，最大固有値が低下する領域が広がった．以上より，既存の競争の減少を抑えることと新たな競争を生み

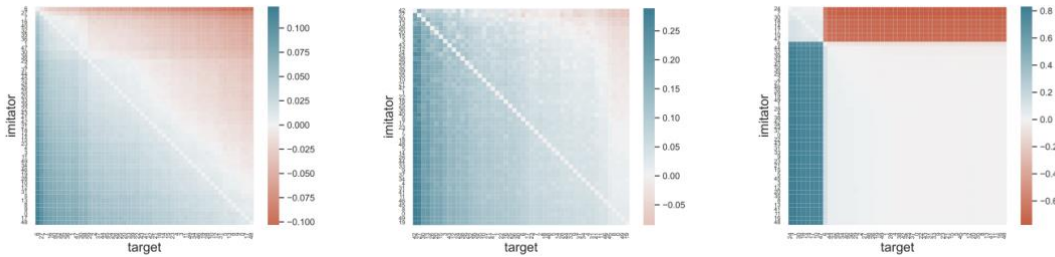


図 1 模倣戦略を適用した際の最大固有値変化 ((左) 疎なネットワーク, (中央) 密なネットワーク, (右) 一部に密な領域を持つネットワーク)

出すことの双方が模倣戦略の成功に重要であることが分かった。

3 複数の不適切発信者を対象とする最適模倣戦略

本章では、前章で述べた手法に関する課題を解決するために提案した最適模倣戦略について述べる。前章の単一ターゲットを対象とする戦略を以降では単純模倣戦略と呼ぶ。

3-1 最適模倣戦略

単純模倣戦略は、競争ネットワーク内で単一の発信者をターゲットとして選び模倣することで全体の注目を抑制する。しかし、この方法はターゲットだけでなく他の発信者との競争関係も変化させるため、一般発信者の注目度を低下させるおそれがある。そこで提案手法では元のモデルにおける競争関係に着目し、ターゲットとの競争のみを選択的に強めることで特定のターゲットへの注目を抑制することを目指す。具体的には、まず不適切コンテンツの発信者を一般の発信者と区別し、不適切発信者との競争を強める一方、一般発信者との不要な競争を低く抑える。

模倣者ノードの特徴ベクトルを c_i 、ターゲットノードの集合を T （その特徴ベクトルを c_τ ）、一般ノードの集合を G （その特徴ベクトルを c_g ）とする。最適な模倣ベクトルを求める問題は、次の多目的最適化問題として定式化される。

$$\begin{aligned} \max_{c_i} \quad & c_i^\top c_\tau \quad (\tau \in T) \\ \text{s. t.} \quad & c_i^\top c_g < \delta, \quad \|c\| = 1 \end{aligned}$$

ここで δ は、一般発信者との競争の許容上限を表す。すなわち、複数のターゲットとの競争を最大化しつつ、一般ノードへの平均的な影響を閾値以下に抑えるという問題である。この最適化問題を解くことで、最適な模倣ベクトル c_i^* が得られる。この多目的最適化問題を解くため、遺伝的アルゴリズムに基づく NSGA-II [22] を用いた。多目的最適化問題の解はパレート集合として得られ、複数のパレート最適解から構成される。パレート集合内の解は互いに優劣がなく、目的関数間のトレードオフを表す。ここからひとつの模倣ベクトルを選択することで、具体的な模倣方法が定まる。

3-2 数値実験

提案手法の有効性を評価するために数値実験を行った。実験に用いる競争ネットワークは 30 ノードからなり、各ノードに 100 次元の特徴ベクトルを割り当てた。各ノードには「不適切コンテンツの発信者」、「模倣者」、「一般ノード」のいずれかのラベルを付与し、本実験では、2 ノードを不適切コンテンツのターゲット、1 ノードを模倣者、残りを一般発信者とした。

まず、注目度変化の典型例を、単純模倣戦略との違いに着目して示す。単純模倣戦略では、ターゲットへの注目が低下する一方、一般発信者の平均注目度もわずかに低下しており、一般発信者への意図しない影響が確認された。これに対し最適模倣戦略では、複数のターゲットそれぞれへの注目が低下するとともに、一般発信者の平均注目度はむしろわずかに上昇しており、一般発信者の注目を低下させていないことが確認された。

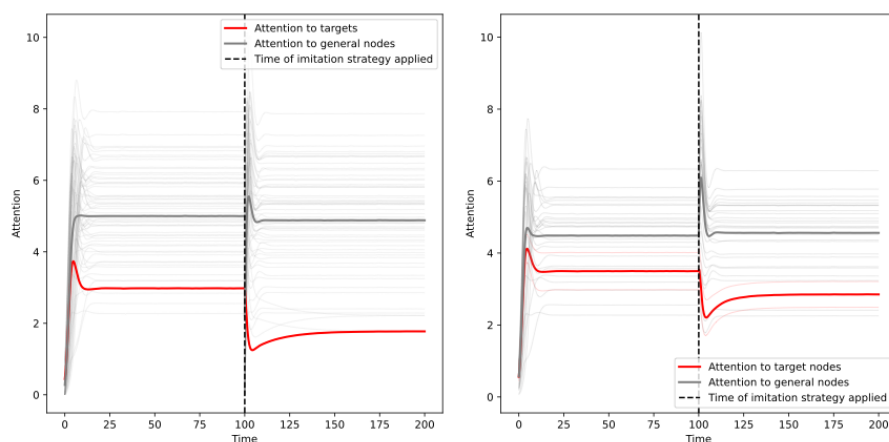


図 2 模倣戦略適用による各ノードの注目度変化（（左）単純模倣戦略、（右）最適模倣戦略）

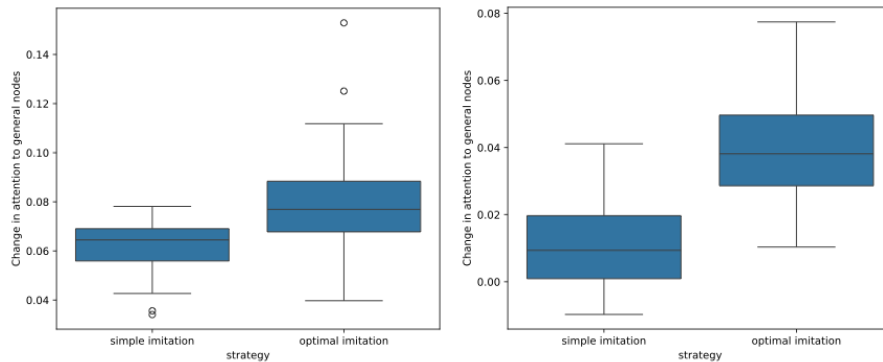


図 3 模倣戦略が一般ノードの注目度に与える影響

次に、模倣戦略が一般ノードへの注目に与える影響を評価した。ランダムに初期化した特徴ベクトルで 50 回の試行を行い、戦略適用前後の一般発信者の平均注目度の差を評価指標とした。この値が負であれば、戦略が一般発信者の注目に負の影響を与えたことを意味する。特徴ベクトルの成分は、相対的に疎なネットワークを生む $[-1,1]$ と、相対的に密なネットワークを生む $[0,1]$ の二種類の範囲から生成した。

疎なネットワークでは、単純・最適いずれの戦略でも一般発信者の平均注目度の低下は見られず、最適模倣戦略の方が平均注目度の増加がやや大きかった。密なネットワークでは、単純模倣戦略の場合に一般発信者の平均注目度が低下する事例が見られたのに対し、最適模倣戦略では低下は見られず、わずかな増加が確認された。これらの結果は、最適模倣戦略が一般発信者の平均注目度の低下を回避するだけでなく、わずかに注目を増加させる効果をもつことを示している。

パレート最適解の間には優劣がないが、どの模倣ベクトルを選ぶかによって注目抑制の効果は異なりうる。目的関数空間におけるパレート最適解の分布には、一方のターゲットとの競争を強めると他方との競争が弱まるというトレードオフが見られた。二つのターゲットをターゲット 1、ターゲット 2 とし、最適模倣ベクトルと各ターゲットの類似度の差を $\Delta = S(c_i^*, c_1) - S(c_i^*, c_2)$ と定義する。すべてのパレート最適解、および模倣者・ターゲットのすべての組合せについて数値シミュレーションを行い、ターゲットへの注目が低下した場合を成功として、 Δ に対する成功率を評価した。ターゲット 1 の成功率は Δ が大きな正の値をとるときに最も高く、 $\Delta = 0$ 付近で最も低くなり、 Δ が負に大きくなると再び上昇した。ターゲット 2 はこれと対称的に、 Δ が大きな負の値で最も高く、 $\Delta = 0$ 付近で最も低かった。すなわち、一方のターゲットとの競争を強めるほど、そのターゲットへの注目を抑制しやすいという直感的な結果が得られた。他方のターゲットへの注目も低下しうるのは、競争を強めると模倣者自身の注目度も低下するため、模倣者が両ターゲットと同時に激しく競争すると自身の影響力が下がり、各ターゲットへの抑制効果が弱まることによると考えられる。

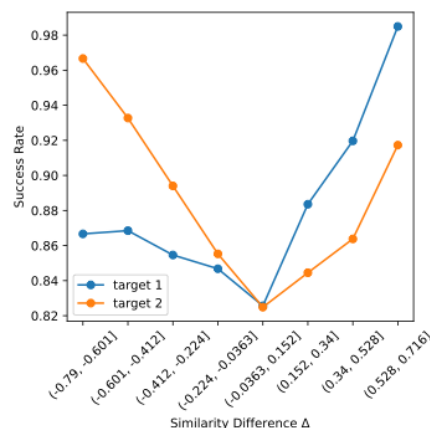


図 4 選択したパレート最適解と成功率の関係

以上の数値実験により、最適模倣戦略は、一般発信者への意図しない悪影響を回避（多くの場合はむしろ微かに注目度が増加）しつつ、複数の不適切発信者への注目を同時に抑制できることが示された。これは、単純模倣戦略が抱えていた一般発信者への影響という課題を解決するものである。さらに、パレート最適解からの模倣ベクトルの選択方法によって抑制効果に変化することを明らかにし、適切な選択により最適模倣戦略の有効性を高められることを示した。

4 おわりに

本研究では、アテンションエコノミーのもとで拡散する不適切コンテンツへの対策として、発信者間の競争に介入する模倣戦略を提案した。今後の課題として、最適化した結果として得られる模倣ベクトルと実際のコンテンツとの対応関係を明確にし、実データを用いた検証が必要である。

【参考文献】

- [1] M. H. Goldhaber, “The attention economy and the net,” *First Monday*, vol. 2, no. 4, 1997.
- [2] T. H. Davenport and J. C. Beck, *The Attention Economy: Understanding the New Currency of Business*. Brighton, MA, USA: Harvard Business Press, 2001.
- [3] V. R. Bhargava and M. Velasquez, “Ethics of the attention economy: The problem of social media addiction,” *Business Ethics Quarterly*, vol. 31, no. 3, pp. 321–359, 2021.
- [4] C. Abidin, “Visibility labour: Engaging with influencers’ fashion brands and #OOTD advertorial campaigns on Instagram,” *Media International Australia*, vol. 161, no. 1, pp. 86–100, 2016.
- [5] D. M. Lazer et al., “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
- [6] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
- [7] J. Pfeffer, T. Zorbach, and K. M. Carley, “Understanding online firestorms: Negative word-of-mouth dynamics in social media networks,” *Journal of Marketing Communications*, vol. 20, no. 1–2, pp. 117–128, 2014.
- [8] M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozaalp, “Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime,” *The British Journal of Criminology*, vol. 60, no. 1, pp. 93–117, 2020.
- [9] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, “Political polarization on Twitter,” in *Proc. International AAAI Conference on Web and Social Media*, vol. 5, no. 1, pp. 89–96, 2011.
- [10] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2003, pp. 137–146.
- [11] Q. Yao, R. Shi, C. Zhou, P. Wang, and L. Guo, “Topic-aware social influence minimization,” in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 139–140.
- [12] M. Kimura, K. Saito, and H. Motoda, “Minimizing the spread of contamination by blocking links in a network,” in *AAAI*, vol. 8, 2008, pp. 1175–1180.
- [13] O. L. Haimson, D. Delmonaco, P. Nie, and A. Wegner, “Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas,” *Proc. ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–35, 2021.
- [14] C. Budak, D. Agrawal, and A. El Abbadi, “Limiting the spread of misinformation in social networks,” in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 665–674.
- [15] P. Lorenz-Spreen, B. M. Mønsted, P. Hövel, and S. Lehmann, “Accelerating dynamics of collective attention,” *Nature Communications*, vol. 10, no. 1, p. 1759, 2019.
- [16] R. M. May and R. H. MacArthur, “Niche overlap as a function of environmental variability,” *Proceedings of the National Academy of Sciences*, vol. 69, no. 5, pp. 1109–1113, 1972.

- [17] M. H. Protter and C. B. Morrey Jr., “Differentiation under the integral sign. Improper integrals. The gamma function,” in *Intermediate Calculus*. Springer, 1985, pp. 421–453.
- [18] E. Laurence, N. Doyon, L. J. Dubé, and P. Desrosiers, “Spectral dimension reduction of complex dynamical networks,” *Physical Review X*, vol. 9, no. 1, p. 011042, 2019.
- [19] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proc. Int. Conf. Machine Learning (PMLR)*, 2014, pp. 1188–1196.
- [20] B. Bamieh, “A tutorial on matrix perturbation theory (using compact matrix notation),” *arXiv preprint arXiv:2002.05001*, 2020.
- [21] P. Bonacich, “Power and centrality: A family of measures,” *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.
- [22] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

〈発 表 資 料〉

題 名	掲載誌・学会名等	発表年月
Method for Mitigating Attention to Inappropriate Content Based on Attention Dynamics Model	IEEE ACCESS	2025 年 4 月
A Strategy to Suppress Attention to Multiple Disseminators of Inappropriate Content	IEEE International Symposium on Autonomous Decentralized Systems	2025 年 7 月
複数の不適切発信者に対する注目度抑制のための最適模倣戦略の提案	電子情報通信学会情報ネットワーク研究会	2025 年 7 月