Vol.2020-CSEC-91 No.17
Vol.2020-SPT-40 No.17
Vol.2020-EIP-90 No.17
2020/11/25

フェイクニュースの伝搬における情報源の信頼度を偽る効果の分析

阿波 拓海[1,a)]　由川 拳都[1]　草野 理沙[1]　市野 将嗣[1,b)]　吉浦 裕[1,c)]

概要：ソーシャルメディアにおけるフェイクニュースの悪影響が深刻になっている．フェイクニュースを信じる人は，同じ価値観の人とだけ交流し，「事実を指摘されるとかえって誤った信念を強める」傾向があるため，フェイクニュースを検知する等の単純な対策では不充分である．一方，ネットワークにおけるユーザの振る舞いと情報伝搬を表現する情報共有モデルの研究が進んでおり，誤った情報の扱いやユーザの非理性的な振る舞いを表現できるため，フェイクニュースの分析・対策への活用が期待できる．しかし，従来の情報共有モデルの研究では，フェイクニュース発信者の持つ意図的な悪意を考慮していない．本論文では，従来の代表的な 3 つの情報共有モデルを取り上げ，悪意の例として，情報源の信用度を実際よりも高く偽る行為をモデルに組み込む．悪意の効果を評価し，情報源の信用度が低いことよりも，低い信用度を高く偽ることの方が大幅な悪影響を与えることを明らかにする．

キーワード：フェイクニュース，情報共有モデル，信用度，マリシャスモデル

1. はじめに

ソーシャルメディア (Facebook，Twitter，Instagram など) が情報共有のための重要な手段となっている．ソーシャルメディアにより，マスメディアが伝えない情報を素早くかつ簡単に手にすることができるようになった．また，個人が簡単に意見を発信できるようにもなった．しかし，意図的な嘘を含む誤情報が大規模に広まりやすくもなった．この誤情報がフェイクニュースと呼ばれ，個人や社会に深刻な影響を与えている．

代表的なフェイクニュース対策として，フェイクニュースの検知 [1–3] と，ニュースの事実確認を行うファクトチェックの自動化 [4,5] がある．しかし，ソーシャルメディアによりエコーチェンバー効果とバックファイヤー効果が促進されるようになった．前者は，価値観が似たもの同士との交流を通して，自分の信念が強くなる心理作用 [6] を指し，後者は事実を指摘されるとかえって誤った信念を強める心理作用を指す [7]．これら 2 つの心理作用により，事実を無視する人々がいるため，フェイクニュースの検知とファクトチェックの自動化が効果的な対策とは限らない．そこで，効果的なフェイクニュース対策を生み出すためには，フェイクニュースがソーシャルメディア内で人々に共有される過程を分析し、フェイクニュースの特性を明らかにする必要がある．

ネットワーク内の情報伝搬を分析するモデルとして，ユーザの振る舞いと情報伝搬を表現する情報共有モデルの研究がマルチエージェントシステムの分野で進んでいる．Glinton らは，友人からの情報に誤りがある状況を想定し，集団内で正しい情報が共有される過程をモデル化した [8]．Tsang らは集団内の多様な意見が少数の主要な意見に収束する過程をモデル化した [9]．しかし，これらのモデルは，フェイクニュース発信者の持つ意図的な悪意を考慮していないため，フェイクニュースの分析・対策に活用するには不十分である．そこで，本研究では，情報共有モデルに悪意に基づく行為を組み込み，その影響を分析する．悪意に基づく行為の例として，情報源の信用度を実際よりも高く偽る行為を取り上げる．

2. 先行研究

2.1 フェイクニュースの特徴分析

Vosoughi らは，フェイクニュースに関する Twitter の投稿を分析し，フェイクニュースは本当のニュースよりも 6 倍早く伝わることを明らかにした [10]．また，彼らはフェイクニュースに関する投稿には驚き，恐れ，嫌悪などを表

[1] 電気通信大学情報理工学研究科
Gradudate School of Informatics and Engineering, The University of Electro-Communications
[a)] taku.awa@uec.ac.jp
[b)] ichino@inf.uec.ac.jp
[c)] yoshiura@uec.ac.jp

す感情的な言葉が含まれることも明らかにした．

Kwon らは，Twitter の投稿とフォローフォロワー関係からフェイクニュースの拡散の性質を分析し，フェイクニュースは，ごく少数の人物が複数回フェイクニュースを流すことで拡散することを明らかにした [11]．また，フェイクニュースに関する投稿には伝聞表現 (例えば,「友人から聞いた話だけど...」) や推量 (例えば,「たぶん」) が含まれることも示した．

さらに，フェイクニュースの拡散には投稿を自動的に共有する「ボット」も使われていることが明らかとなっている [12]．

以上の研究は，フェイクニュースやフェイクニュースに関わる人の特徴を表面的には分析しているが，人がフェイクニュースをどの程度信じているか，フェイクニュースを伝えた人の間にどの程度の信頼関係があるかといった深層の特徴までは分析していない．マルチエージェント分野の情報共有モデルは，深層の特徴までモデル化しているので，情報共有モデルの利用によりフェイクニュースのより深い分析が可能になると期待される．

2.2 フェイクニュースにかかわる情報共有モデル

フェイクニュースの伝搬に関係する情報共有モデルの研究は主に 2 種類に分類できる．第一は，ニュースの真偽を前提としたモデル化である．なかでも Glinton らは，悪意はないが誤った情報が存在する状況下で正しい意見を共有する過程をモデル化している [8]．Glinton らは意見決定をベイズ更新に基づいてモデル化しているが，実際の人間はそのように完全に合理的に判断するとは限らない．そこで，より自然な意見形成方法を取り入れた Kozma らのモデルも取り上げる [13]．

第二のタイプの研究では，ニュースの真偽を前提としないモデル化である．なかでも Tsang らは，自分と似た意見を持つ人の意見を受け入れやすいという人間の性質に基づいて，多様であった意見が集約する過程をモデル化している [14]．以下では上記の 3 つのモデルの概要を述べる．

2.2.1 Glinton モデル

Glinton らは人の意見に誤りが存在する環境下で，集団が正しい意見を共有することを意見共有問題と呼び，この問題をモデル化した [8]．この意見共有問題では，ネットワーク上の各個人（以下，エージェントと呼ぶ）が外部の真実 $b \in \mathrm{True}, \mathrm{False}$ と自分の意見の一致を目標とし，隣接エージェント同士で意見のやり取りを行う．エージェントには外部の真偽情報を直接受け取るセンサーエージェントと，それ以外の一般エージェントの 2 種類が存在する．センサーエージェント $s_j (j = 1, 2, ..., |S|, S$ はセンサーエージェントの集合) は信念値 $P_j(b = \mathrm{True}) \in [0, 1]$ を持つ．これは s_j が $b = \mathrm{True}$ と信じる主観確率である．$b = \mathrm{False}$ の主観確率は $P_j(b = \mathrm{False}) = 1 - P_j(b = \mathrm{True})$ となる．各ステップ k ごとにセンサーエージェント s_j はセンサーから真偽の値を受け取り，以下のベイズ更新の式によって信念値を更新する．

$$P_j^k(b = \mathrm{True}) = \frac{rP_j^{k-1}(b = \mathrm{True})}{rP_j^{k-1}(b = \mathrm{True}) + (1 - r)(1 - P_j^{k-1}(b = \mathrm{True}))} \quad (1)$$

ここで, r はセンサーが正しい意見を伝達してくる確率であり, センサーからの意見 o_s をとすると $r = P(o_s = \mathrm{True}|b = \mathrm{True})$ と表す．意見を決定する閾値を $\sigma(0.5 < \sigma < 1.0)$ とするとき，更新後の信念値が $P^k > \sigma\ (< 1 - \sigma)$ となった場合，意見 o_i を True(False) に決定する．意見が決定された場合，隣接するエージェントにその意見を発信する．

一般エージェント $a_i (i = 1, 2, ..., |A|, A$ は一般エージェントの集合) は，センサーエージェント同様に信念値 $P_i(b = \mathrm{True}) \in [0, 1]$ を持つ．隣接エージェント j から意見を受け取る時，以下のベイズ更新の式によって信念値を更新する．

$$P_i^k(b = \mathrm{True}) = \frac{cpP_i^{k-1}(b = \mathrm{True})}{cpP_i^{k-1}(b = \mathrm{True}) + cp'(1 - P_i^{k-1}(b = \mathrm{True}))} \quad (2)$$

ここで $cp = P(o_j = \mathrm{True}|b = \mathrm{True})$ であり，ニュースの真実が True であるときにエージェント j から True を受け取る確率である．$cp' = P(o_j = \mathrm{True}|b = \mathrm{False})$ であり，ニュースの真実が False であるときにエージェント j から True を受け取る確率である．そしてセンサーエージェント同様に閾値 σ に基づいて意見を決定する．

以上の手順をネットワーク内のすべてのエージェントが繰り返し行うことで意見が伝搬する．

2.2.2 Kozma モデル

Kozma らは，Glinton モデルと同じ条件下において，異なる信念値更新方式を提案した [13]．エージェント a_i が隣人 a_j から意見を受け取ると，Glinton 方式ではベイズ更新の式を用いて信念値の更新を行っていたが，この方式ではより単純な以下の重み平均の式により信念値を更新する．

$$P_i^k(b = \mathrm{True}) = (1 - t)P_i^{k-1}(b = \mathrm{True}) + tP_j^{k-1}(b = \mathrm{True}) \quad (3)$$

ここで $t \in [0, 1]$ が他人の意見の信頼度であり，これが大きいほど他人の意見を受け入れやすくなる．センサーエージェントの場合，t がセンサーが正しい情報を伝達してくる確率であり，センサーの伝達する意見が True/False の場合に $P_j^{k-1}(b = \mathrm{True}) = 1/0$ とする．

2.2.3 Tsang モデル

Tsang らはネットワーク上に多様な意見を持つエージェントが存在する環境において，意見がどのように収束するかをモデル化した [9]．このモデルの各エージェントは，自

Vol.2020-CSEC-91 No.17
Vol.2020-SPT-40 No.17
Vol.2020-EIP-90 No.17
2020/11/25

分と考えが似ている人の意見は受け入れやすく，異なる考えを持つ人の意見は受け入れにくいという性質を持つ．

各エージェントは自分の意見を示す値である意見値 $x_i (x_i \in [0,1])$，隣接エージェント j からの意見の重み $w_{i,j}$ を持つ．この意見値と重みは以下の式を用いて更新を行う．

$$x_i \leftarrow \frac{w_{i,i}x_i + \sum_{j\in N(i)} w_{i,j}x_j}{w_{i,i} + \sum_{j\in N(i)} w_{i,j}} \tag{4}$$

$$w_{i,j} \leftarrow \frac{w_{i,j} + rT(x_i, x_j)}{1+l} \tag{5}$$

ここで $N(i)$ はエージェント i とリンクを持つノードの集合である．l は学習率といい，この値が大きければより自分と異なる意見を受け入れにくくなる．また，$T(i)$ は他のエージェントの意見の信用関数であり，以下で表す．

$$T(x, x') = \exp\left(-\frac{(x-x')^2}{h}\right) \tag{6}$$

h は意見への共感度を表し，この値が大きければより他人の意見を受け入れやすくなる．

エージェントの意見値の初期値を，$x_i \in [0,1]$ の一様乱数とするエージェントが全体の 80 %，$x_i = 0$ または $x_i = 1$ の極端な意見で初期化するエージェントがそれぞれ全体の 10 %とする．極端な意見を持つエージェントは意見値や重みの更新は行わない．各エージェントの重みの初期値は，エージェント i, j の次数 d_i, d_j を用いて以下のように初期化する．

$$w_{i,j} = \frac{d_j}{d_i}, \quad w_{i,i} = \frac{d_i}{d_i} = 1 \tag{7}$$

ここで，d_i はエージェント i とリンクを持つノードの数である．つまり，リンクが多いほどその人の意見は信頼できるということになる．意見値と重みの更新は全エージェントの意見値の変化が閾値以下になるまで行う．

3. 研究方針

フェイクニュースの拡散においては悪意あるエージェントが，他のエージェントに自分の意見を信用させ，誤った意見を拡散させることが想定される．しかし，先行研究の意見伝搬モデルは，いずれも，悪意のあるエージェントの存在を想定していなかった．

そこで，2.2 節で述べた各意見伝搬モデルにおいて，誤った意見を拡散させようとする悪意のあるエージェントを導入し，その影響を調べる．具体的には悪意のあるエージェントは，自分の信用度を実際の値よりも大きな値に見せかけ，自分の意見の信頼性を高くして意見へと誘導しようとする場合を検討する．4 章では Glinton のモデル，5 章では Kozma のモデル，6 章では Tsang のモデルにおいて上記の信用度の詐称の効果を分析する．

4. Glinton モデルにおける不正操作の影響

2.2.1 節で述べた Glinton モデルにおいて情報の発信源となるセンサーの信頼度を不正操作する場合の影響を評価し，結果について考察する．

4.1 信用度の不正操作方法

センサーは自分の信頼度が本来は r であるところを $r' = 1 - r$ であるとセンサーエージェントに見せかけるという不正操作を行う．例えば，$r = 0.1$（センサーエージェントがセンサーから受け取る意見の 10 %が正しく 90 %が誤り）を $r = 0.9$（90 %が正しい）に見せかける．

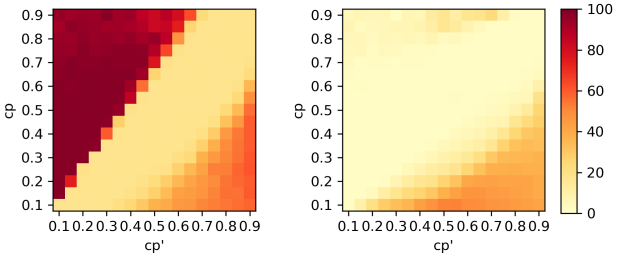
意見伝搬を行うためのネットワーク構造として，スケールフリーネットワーク [15] を用いた．これは，ごく一部のノードが多くのノードとリンクをもち，大多数のノードは少数のノードとしかリンクをもたないという状況を表現したネットワークであり，ソーシャルメディア，航空網，学術論文の共著関係など世の中の多くのネットワーク構造を表現していることが知られている．そこで，このスケールフリーネットワークを用いることで，ソーシャルメディア上で悪意のある人物が情報源を不正操作し，フェイクニュースを拡散させようとする状況を分析する．

今回はニュースの真実 b = False，全エージェント数（ノード数）を 100 とし，そのうちセンサーエージェントが 20 とした．そして，全エージェントの信念値の初期値は $P^0 = 0.5$ とした．一般エージェントは同じ信頼度 cp と不信度 cp' の組 (cp, cp') を持つことにする．これは 0.1 から 0.9 まで 0.05 で刻んだ値とし，全ての (cp, cp') の組み合わせ (0.1,0.1),(0.1,0.15),...,(0.9,0.9) に対して 100 回ずつシミュレーションを実行した．1 回のシミュレーションの終了条件は，全員の意見が決定するまたは信念値の更新回数が K 回を超える（今回は $K = 1000$ とした）までとした．最後にシミュレーション終了時の，真実 b = False と同じ意見 o = False を持つエージェントの割合（以下，正解率と呼ぶ）と真実とは反対の意見 o' = True を持つエージェントの割合（以下，不正解率と呼ぶ）を求めた．また，意見決定の閾値を $\sigma = 0.9$ とした．センサーの信頼度は $r = 0.1, 0.9$ の 2 パターンとし，信頼度を不正操作したときとしていないときで結果を比較した．

4.2 センサーの信頼度を不正操作しない場合の実験評価

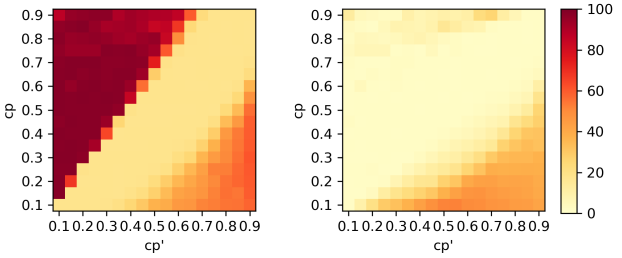
センサーの信頼度 $r = 0.1$ および $r = 0.9$ で不正操作していない場合の評価結果を示す（図 1，図 2）． 各図において，図中の (a) 正解率の場合は色が赤色に近いほど真実と同じ意見 (False) を持つエージェントが多いことを示す．(b) 不正解率の場合は色が赤色に近いほど真実と異なる意見 (True) を持つエージェントが多いことを示す．また，cp, cp' の組が cp が高く cp' が低い（グラフ上では左上に近い）ほど一般エージェントは他人の意見を信頼しやすく，cp が低く cp' が高い（グラフ上では右下に近い）ほど信頼しにくい．

Vol.2020-CSEC-91 No.17
Vol.2020-SPT-40 No.17
Vol.2020-EIP-90 No.17
2020/11/25



(a) 正解率　　**(b)** 不正解率

図 1: $r = 0.1$，センサーの信頼度の不正操作なしの場合



(a) 正解率　　**(b)** 不正解率

図 2: $r = 0.9$，センサーの信頼度の不正操作なしの場合

図 1，図 2 のどちらにおいても，cp が高く cp' が低いほど正解率が高い，cp が低く cp' が高いほど不正解率が高いという結果が得られた．また，グラフの形状がほぼ同一であり，センサーの信頼度がどちらの場合も正解率，不正解率はほぼ等しいと言える．このような結果が得られた理由について考察する．今回はニュースの真実を $b = \text{False}$ としている．

最初に $r = 0.1$ の場合を考える．この時センサーエージェントは確率 0.9 でセンサーから True の意見（不正解の意見）を受け取る．センサーエージェントは以下の更新式により，意見を更新する．

$$P^k(b = \text{True}) = \frac{r(1 - P^{k-1}(b = \text{True}))}{r(1 - P^{k-1}(b = \text{True})) + (1 - r)P^{k-1}(b = \text{True})} \quad (8)$$

$r = 0.1, P^k(b = \text{True}) = 0.5$ を代入すると，$P^k(b = \text{True}) = 0.1$ となり，s_i の意見は False（正解の意見）と決定する．図 1(a) の左上部分では，一般エージェントは隣人を信用するので，False がそのまま伝搬するため，正解率が高くなる．図 1(b) の右下部分では，一般エージェントは隣人を信用しないので，一般エージェントの意見更新の際に False が True に反転するため，不正解率が高くなる．センサーエージェントは確率 0.1 でセンサーから False（正解）の意見を受け取る．この場合は，図 1 の正解率，不正解率は上記の True（不正解）の意見を受け取った場合と逆になる．しかし，False の意見を受け取る確率は小さいため，全体として，図 1 のようになる．

次に $r = 0.9$ の場合を考える．この時センサーエージェントは確率 0.1 でセンサーから True の意見（不正解の意見）を受け取る．センサーエージェントは以下の更新式により，意見を更新する．

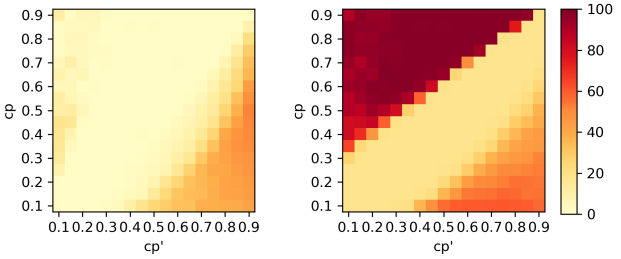
$$P^k(b = \text{True}) = \frac{(1 - r)P^{k-1}(b = \text{True})}{(1 - r)P^{k-1}(b = \text{True}) + r(1 - P^{k-1}(b = \text{True}))} \quad (9)$$

$r = 0.9, P^k(b = \text{True}) = 0.5$ を代入すると，$P^k(b = \text{True}) = 0.1$ となり，s_i の意見は False（正解の意見）と決定する．よって一般エージェントの間での意見伝搬は，図 1 の場合と同様になり，全体として図 2 のようになる．

以上の考察をまとめると，センサーの信頼度を正しく認識していれば，センサーが誤った情報を伝達してもセンサーエージェントによってそれが正しい意見に修正され，ネットワーク内に正しい情報を伝搬することができる．また，センサーの信頼度が高い場合は，センサーが正しい情報を伝達したときセンサーエージェントはそれを正しいと認識してネットワーク内に正しい情報を伝搬することができるといえる．

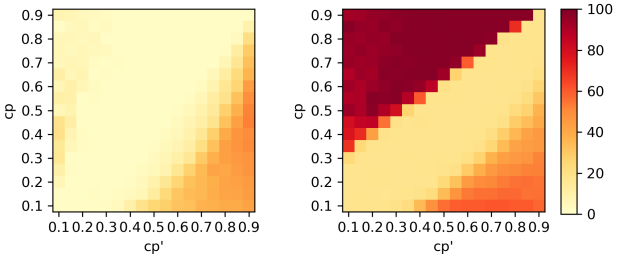
4.3 センサーの信頼度を不正操作した場合の実験評価

センサーの信頼度が本来 $r = 0.1$ および $r = 0.9$ であるところを，センサーエージェントには $r' = 0.9$，$r' = 0.1$ であるように見せかけるという不正操作を行った場合の評価結果を示す（図 3，図 4）．



(a) 正解率　　**(b)** 不正解率

図 3: $r = 0.1$，センサーの信頼度の不正操作ありの場合



(a) 正解率　　**(b)** 不正解率

図 4: $r = 0.9$，センサーの信頼度の不正操作ありの場合

図 3(a)、図 4(a) に示すように，cp が高く cp' が低いほど

（友人を信用するほど）正解率が低い．また、図 3(b)、図 4(b) に示すように，友人を信用するほど不正解率が高い．図 1，図 2 の場合と同様にグラフの形状がほぼ同一であり，センサーの信頼度がどちらの場合も正解率，不正解率はほぼ等しいと言える．このような結果が得られた理由について考察する．

最初に，$r = 0.1$ の場合について考える．不正操作が行われた場合，センサーエージェント s_i は (8) 式の r を r' に置き換えた以下の式で信念値更新を実行する．

$$P^k(b = \text{True}) = \frac{r'(1 - P^{k-1}(b = \text{True}))}{r'(1 - P^{k-1}(b = \text{True})) + (1 - r')P^{k-1}(b = \text{True})} \quad (10)$$

$r' = 0.9, P^k(b = \text{True}) = 0.5$ を代入すると，$P^k(b = \text{True}) = 0.9$ となり，s_i の意見は True（すなわち不正解の意見）と決定する．図 3(a) の左上部分では，一般エージェントは隣人を信用するので，True がそのまま伝搬し，正解率が低くなる．図 3(a) の右下部分では，一般エージェントは隣人を信用しないので，True が False に反転し，正解率がやや高くなる．図 3(b) の左上部分では、一般エージェントが隣人を信用するので，True がそのまま伝搬し，不正解率が高くなる．

次に，$r = 0.9$ の場合について考える．不正操作が行われた場合，センサーエージェント s_i は (9) 式の r を $r' = 1 - r$ に置き換えた以下の式で信念値更新を実行する．

$$P^k(b = \text{True}) = \frac{((1 - r')P^{k-1}(b = \text{True})}{(1 - r')P^{k-1}(b = \text{True}) + r'(1 - P^{k-1}(b = \text{True}))} \quad (11)$$

$r' = 0.1, P^k(b = \text{True}) = 0.5$ を代入すると，$P^k(b = \text{True}) = 0.9$ となり，s_i の意見は True（すなわち不正解の意見）と決定する．よって一般エージェントの間での意見伝搬は，図 3 の場合と同様になる．

以上の考察をまとめると，センサーの信頼度を本来より高く誤認している場合には，センサーが発信した誤った情報をセンサーエージェントは正しいと認識し，ネットワーク内にフェイクニュースが伝搬する．また，センサーの信頼度を本来より低く誤認している場合には，センサーが発信した正しい情報がセンサーエージェントによってそれが誤った意見に修正され，ネットワーク内にフェイクニュースが伝搬すると言える．

5. Kozma モデルにおける不正操作の影響

Kozma モデルにおいて，情報の発信源となるセンサーの信頼度を不正操作した場合の影響を評価する．Glinton モデルとの違いは，第一に信念値の更新式がベイズ更新の式から単純な重み平均の式へと簡略化されている点である．また，不信度 cp' に相当するパラメータを持たず，他人の意見の重みである信用度 t のみを持つ．この方式において 4 章と同じパラメータや条件を用いて実験を行った．t は全員共通の値で，0.1 から 0.9 まで 0.05 で刻み，それぞれの値について 100 回ずつシミュレーションを実行した．正解率と不正解率，そしてどちらにも意見を決定しなかったエージェント数（以下，意見未決定率と呼ぶ）を求めた．

5.1 センサーの信頼度を不正操作しない場合の実験評価

センサーの信頼度 $r = 0.1$ および $r = 0.9$ で不正操作しない場合の評価結果を示す（図 5，図 6）． 各図において，

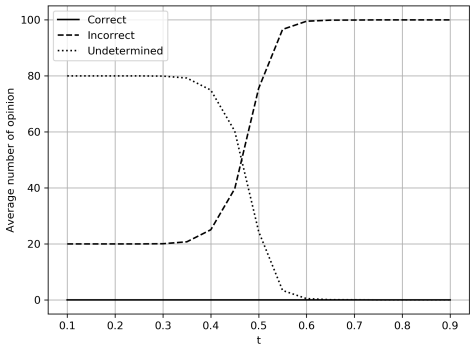


図 5: $r = 0.1$，センサーの信頼度の不正操作なしの場合

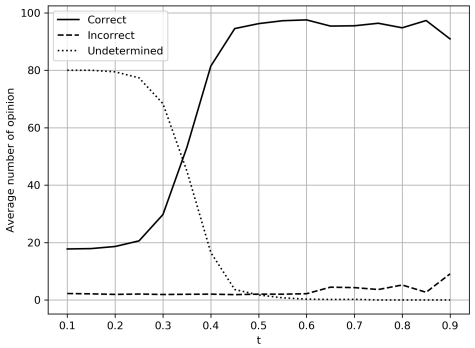


図 6: $r = 0.9$，センサーの信頼度の不正操作なしの場合

実線が正解率，破線が不正解率，点線が意見未決定率の平均を表す．

図 5 においては，正解したエージェントはほぼ存在せず，一般エージェントの信頼度が $t = 0.4$ を越えると急激に誤った意見が広まることがわかる．図 6 では，正解も不正解も存在するが，$t = 0.3$ を越えると急激に正しい意見が広まることが読み取れる．

このような結果が得られた理由について考察する．センサーエージェント i は (3) 式に $t = r$ を代入した以下の式で更新を行う．

$$P_i^k(b = \text{True}) = (1 - r)P_i^{k-1}(b_{s_i}) + rP_j^{k-1}(b_{a_j}) \quad (12)$$

図 5 の場合は，真実は $b = \text{False}$ であり，センサーの信頼度は $r = 0.1$ であるため，センサーからは 0.9 の確率で True

の意見（不正解の意見）が伝達する．センサーエージェントは (12) 式を用いて更新する．(12) 式は重み平均なので，意見を数回受け取ると一定数のセンサーエージェントは意見を True へと決定する．その意見は一般エージェントにも伝達するが，信用度 t が低い状態では更新後も意見が決まらず，ほとんどのエージェントが意見未決定である．しかし，信用度が $t = 0.4$ を越えるとその意見 True（不正解の意見）がネットワーク全体へ伝搬する．

図 6 の場合は反対に，センサーからは 0.9 の確率で False の意見（正解の意見）が伝達し，センサーエージェントが (12) 式を用いて更新することで意見を False へと決定する．一般エージェント間での意見伝達は図 5 の場合と同様に，信用度が小さいと意見が決定しにくく，信用度が $t = 0.3$ を越えると False の意見が拡散すると考える．一方で，センサーから 0.1 の確率で True の意見が伝達し，一部のセンサーエージェントは意見を True に決定するため，信用度の増加とともにこの意見も少数ではあるがネットワーク内に拡散したと考える．

以下，式を用いて説明する．図 5 の場合は，センサーエージェントはセンサーが伝達する真偽の値を $o_s \in \{\text{True}, \text{False}\}$ とするとき，(12) 式に $t = r = 0.1$ を代入した以下の式で更新を行う．

$$P_i^k = 0.9P_i^{k-1} + 0.1 \times 1 \qquad (when\ o_s = \text{True}) \quad (13)$$

$$P_i^k = 0.9P_i^{k-1} + 0.1 \times 0 \qquad (when\ o_s = \text{False}) \quad (14)$$

$r = 0.1$ よりセンサーからは 90 %の確率で $o_s = \text{True}$（不正解意見）が伝達するため，(13) 式が支配的である．センサーエージェントは信念値が $P_i^k \geq 0.55$ となり，意見決定には至らないが，センサーからの意見を複数回入力すると一定数のセンサーエージェントが意見を True に決定し発信する．一般エージェントはその意見を受け取り，(3) 式に基づいて更新を行う．このとき，t が大きくなるほど他人の意見を受け入れやすくなり，$t = 0.4$ 以上でネットワーク内に True の意見が広まる．

図 6 の場合は，(12) 式に $t = r = 0.1$ を代入した以下の式で更新を行う．

$$P_i^k = 0.1P_i^{k-1} + 0.9 \times 1 \qquad (when\ o_s = \text{True}) \quad (15)$$

$$P_i^k = 0.1P_i^{k-1} + 0.9 \times 0 \qquad (when\ o_s = \text{False}) \quad (16)$$

$r = 0.9$ より (16) 式が支配的であり，大多数のセンサーエージェントはこの式で信念値を更新し，False（正解意見）へと決定する．その結果，図 5 の場合と同様にして，ネットワーク内でも False の意見が拡散され，正解率が高くなったと考える．一方で，一部のセンサーエージェントは (15) 式により，意見を True に決定する．この影響により，約 1 割のセンサーエージェントとそれに隣接する一般エージェントの間では True の意見が流れ，t が大きくなるにつれてグラフ上で不正解の割合が増加したと考える．

以上をまとめると，センサーの信頼度に偽りがない場合，センサーの信頼度が低いとセンサーは高い確率で誤るが，センサーの重みが小さいため，センサーエージェントが誤った意見を持つことが抑止される．センサーの信頼度が高いとセンサーは高い確率で正しい意見を伝達し，センサーの重みが大きいため，センサーエージェントは正しい意見を迅速に決定する．

5.2 センサーの信頼度を不正操作した場合の実験評価

センサーの信頼度が本来 $r = 0.1$ および $r = 0.9$ であるところを，センサーエージェントには $r' = 0.9$，$r' = 0.1$ であるように見せかけるという不正操作を行った場合の評価結果を示す（図 7，図 8）．

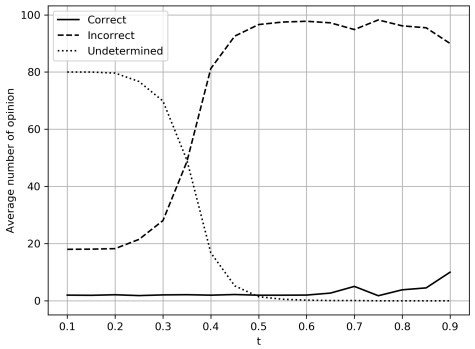


図 7: $r = 0.1$，センサーの信頼度の不正操作ありの場合

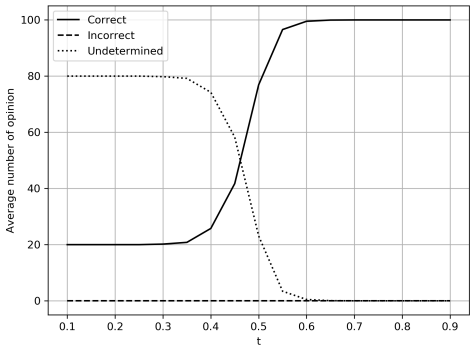


図 8: $r = 0.9$，センサーの信頼度の不正操作ありの場合

図 7 においては，図 5 と比較して t が小さい時点で誤った意見がネットワーク内で急速に拡散していることが分かった．一方，図 8 では，図 5 と比較して t が大きくならないと，正しい意見が伝搬しにくいことが分かった．その理由について考察する．

図 7 の場合，センサーエージェントは (3) 式に $t = r' = 0.9$ を代入したもので信念値を更新する．これは (15) 式と (16) 式と等しいが，実際には $r = 0.1$ であるため，(15) 式が支配的である．センサーからは 90 %の確率で True の意見（不正解意見）が伝達するため，センサーエージェントは (15) 式で更新を行って，True の意見が拡散する．センサーの

信頼度を偽らなかった場合の (13) 式による意見更新よりもセンサーの不正解意見の重みが大きいので，より多くのエージェントが不正解意見を持つ．その結果，図 5 よりも t が小さい時点で不正解意見が拡散する．

同様に，図 8 の場合のセンサーエージェントの信念値更新式は，(3) 式に $t = r' = 0.1$ を代入したものである．センサーからは 90 %の確率で False の意見（正解意見）が伝達するが，(14) 式ではセンサーの重みが小さいので，正解意見が広まりにくい．そのため，t が大きくないならないと正解意見が伝搬しにくいと考えられる．

以上をまとめると，センサーの信頼度に偽りがある場合には，センサーの信頼度が低いとセンサーは高い確率で誤り，センサーの重みが大きいため，センサーエージェントが誤った意見を迅速に決定する．センサーの信頼度が高いとセンサーは高い確率で正しい意見を伝達するが，センサーの重みが小さいため，センサーエージェントが正しい意見を持つことが抑止される．

6. Tsang モデルにおける不正操作の影響

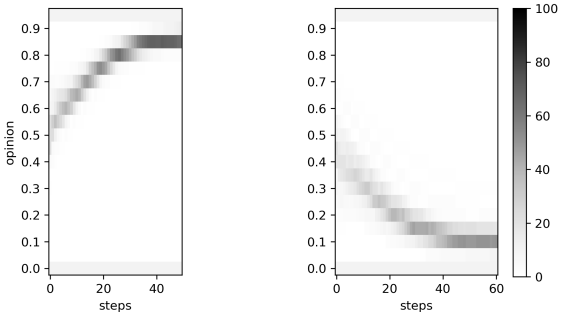
6.1 信用度の不正操作方法

Tsang 方式では極端な意見を持ちかつ重みが大きい人に，意見が吸い寄せられていくことで徐々に意見が一極化する．そこで，本稿では一方の極端な意見を持つエージェントの重みを 8 倍に見せかけるという不正操作を行う．重みはエージェントの次数（直接の友人数）に比例するので，友人が多いように見せかける，あるいは，ボットのエージェントを多数作成して友人とすることにより，重みを大きくすることができる．全エージェント数（ノード数）を 100 とし，そのうち意見値の初期値 $x_i^0 = 0$ および $x_i^0 = 1$ の極端な意見を持つエージェントをそれぞれ全体の 10 %とした．そのほかの 80 %は，$x_i \in [0, 1]$ の一様乱数で初期化を行った．極端な意見を持つエージェントの数の偏りによって生じる影響と比較するため，$x_i^0 = 0$ を 10 %，$x_i^0 = 1$ を 5 %，一様乱数で初期化するものを 85 %とした実験も行った．また，各エージェントの重みを各々の次数によって初期化した．ステップ毎の 100 エージェントの意見値の分布を調べた．1 回のシミュレーションの終了条件は，全員の信念値の変化が閾値未満 ($\delta < 0.001$) となるまでとし，これを 10 回繰り返した．重みの不正操作をしたときとしていないときで結果を比較した．

6.2 センサーの信頼度を不正操作しない場合の実験評価

重みの不正操作をしていない場合の評価結果の例を示す（図 9，図 10）．

図の横軸は更新回数であり，縦軸は上に近いほど 1 の意見に近く，下に近いほど 0 の意見に近いことを表す．更新回数が 0 の時（初期値の時）には，意見 0 と 1 のエージェントが各々 10 %存在する以外は，意見が一様分布になっている．その後，意見分布が 0 または 1 に集約していく．



(a) 1 への収束の例　　**(b)** 0 への収束の例

図 9: 不正操作なし,$x_i = 0$ が 10 %,$x_i = 1$ が 10 %の場合

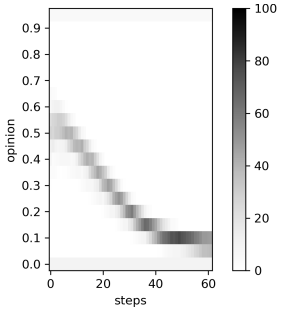


図 10: 不正操作なし,$x_i = 0$ が 10 %,$x_i = 1$ が 5 %の場合

図 9 の場合は，全体として 0 に近い意見へ収束する場合と 1 に近い意見へ収束する場合がそれぞれ 5 回ずつ見られた．0 に収束した一例を図 9(a) に，1 に収束した一例を図 9(b) に示す．この例では，0 への収束の方が 1 への収束よりも速かったが，10 回の実験全体としては，0 への収束と 1 への収束の速さに有意差はなかった．極端な意見を持つエージェントが同数であるため，ネットワーク構造の違いによって意見の収束方向が変化したが，確率的なものであると考える．

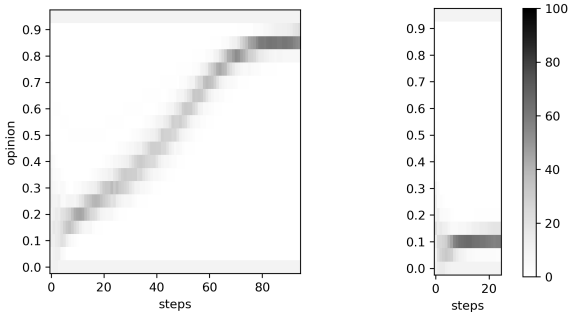
次に図 10 の場合は，10 回全てにおいて全体が 0 に近い意見へと収束した．これは $x_i = 0$ の極端な意見を持つエージェントの方がネットワーク内に多く存在するため影響が大きいためである．

6.3 センサーの信頼度を不正操作した場合の実験評価

$x_i = 0$ という極端な意見を持つエージェントが，重みを 8 倍に見せかけるという不正操作を行った場合の評価結果の例を示す（図 3，図 4）．

図 11 の場合は，0 に近い意見へ収束する場合と 1 に近い意見へ収束する場合がそれぞれ 5 回ずつ見られた点は図 9 と同様であった．一方で 0 への収束の平均ステップ数は 101.3 であり，1 への収束の平均ステップ数は 30.8 であった．これは 0 の意見を持つエージェントの重みが不正に大きくなっているため 1 への意見の収束が速くなっている．

図 12 の場合は，図 10 の場合と同様の理由に 10 回全てにおいて全体の意見が 0 の方向へと収束した．しかし，図



(a) 1 への収束の例　　**(b)** 0 への収束の例

図 11: 不正操作あり,$x_i = 0$ が 10 %,$x_i = 1$ が 10 %

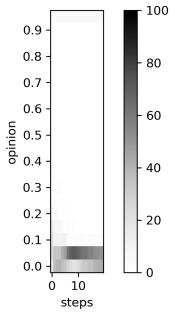


図 12: 不正操作あり,$x_i = 0$ が 10 %,$x_i = 1$ が 5 %

10 の場合と比較して，収束までの平均ステップ数が 44.8 から 20.8 まで短くなっていた．

以上の結果をまとめると，重みを大きく見せかけるという不正操作を行うことで，全体の意見を自らの方へより早く誘導したり，反対の意見への収束を遅くするといったことが可能であると言える．

7. おわりに

情報共有モデルをフェイクニュースの分析・対策に活用することを目的として、代表的な 3 つの情報共有モデルを取り上げ，情報源やエージェントの信頼度を実際より低く偽ることを評価した．

ニュースの真偽を前提とし，情報源の信頼度に応じてベイズ更新によりニュースを伝搬する Glinton らのモデルの場合，情報源の信頼度が低くても，エージェントが信頼度を正しく把握していれば，誤った情報が訂正されるため，ネットワーク全体に誤った情報が拡散することはなかった．しかし，信頼度の低い情報源について信頼度を高く偽ると，誤った情報がそのままネットワーク全体に拡散した．

ベイズ更新の代わりに重み平均によりニュースを伝搬する Kozma らのモデルの場合，情報源の信頼度が低くても，エージェントが信頼度を正しく把握していれば，エージェントが誤った情報の影響を受けにくいため，ネットワーク全体への誤情報の拡散を遅くすることができた．しかし，信頼度の低い情報源について信頼度を高く偽ると，誤った情報の拡散が抑止されなかった．

ニュースの真偽を前提とせず，多様な意見から少数の意見への収束を表現する Tsang らのモデルの場合，情報源の信頼度を実際よりも大きく偽ることで，全体の意見を当該情報源の意見に向けて，早く収束させることができた．

以上により，情報源の信用度が低いことよりも，低い信用度を高く偽る方が大幅な悪影響を与えることを明らかにした．

参考文献

[1] Wang, Y. et al.: Eann: Event adversarial neural networks for multi-modal fake news detection, *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pp. 849–857 (2018).
[2] Tacchini, E. et al.: Some like it hoax: Automated fake news detection in social networks, *arXiv preprint arXiv:1704.07506* (2017).
[3] Ma, J., Gao, W. and Wong, K.-F.: Rumor detection on twitter with tree-structured recursive neural networks, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, Association for Computational Linguistics (2018).
[4] Hassan, N., Arslan, F., Li, C. and Tremayne, M.: Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster, *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1803–1812 (2017).
[5] Vo, N. and Lee, K.: The rise of guardians: Fact-checking url recommendation to combat fake news, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 275–284 (2018).
[6] Quattrociocchi, W. et al.: Echo chambers on Facebook, *Available at SSRN 2795110* (2016).
[7] Nyhan, B. and Reifler, J.: When corrections fail: The persistence of political misperceptions, *Political Behavior*, Vol. 32, No. 2, pp. 303–330 (2010).
[8] Glinton, R. T. et al.: Towards the understanding of information dynamics in large scale networked systems, *Information Fusion, 2009. FUSION'09. 12th International Conference on*, IEEE, pp. 794–801 (2009).
[9] Tsang, A. and Larson, K.: Opinion dynamics of skeptical agents, *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 277–284 (2014).
[10] Vosoughi, S. et al.: The spread of true and false news online, *Science*, Vol. 359, No. 6380, pp. 1146–1151 (2018).
[11] Kwon, S., Cha, M. and Jung, K.: Rumor detection over varying time windows, *PloS one*, Vol. 12, No. 1, p. e0168344 (2017).
[12] Ferrara, E. et al.: The rise of social bots, *Communications of the ACM*, Vol. 59, No. 7, pp. 96–104 (2016).
[13] Kozma, B. and Barrat, A.: Consensus formation on coevolving networks: groups' formation and structure, *Journal of Physics A: Mathematical and Theoretical*, Vol. 41, No. 22, p. 224020 (2008).
[14] McPherson, M., Smith-Lovin, L. and Cook, J. M.: Birds of a feather: Homophily in social networks, *Annual review of sociology*, Vol. 27, No. 1, pp. 415–444 (2001).
[15] Barabási, A.-L. and Albert, R.: Emergence of scaling in random networks, *science*, Vol. 286, No. 5439, pp. 509–512 (1999).