DEIM2021 H33-5

Twitterにおけるフェイクニュース拡散モデルの提案

村山　太一† 　若宮　翔子† 　荒牧　英治† 　小林　亮太††,†††

† 奈良先端科学技術大学院大学 〒 630—0192 奈良県生駒市高山町 8916-5
†† 東京大学 〒 113–8654 東京都文京区本郷 7-3-1
††† 科学技術振興機構 さきがけ
E-mail: †{murayama.taichi.mk1,wakamiya,aramaki}@is.naist.jp, ††r-koba@k.u-tokyo.ac.jp

あらまし　近年，モバイルデバイスやインターネットの普及により，人々が様々なフェイクニュースを目にするリスクが高まっている．このような背景からも，フェイクニュースがオンラインでどのように拡散するのかを理解し，それを説明できる数理モデルの開発が不可欠である．本研究では，Twitter 上におけるフェイクニュースの拡散を説明する新たな点過程モデルを提案する．提案モデルでは，フェイクニュースの拡散は 2 段階のプロセスとして説明される．1 段階目では，フェイクニュースが普通のニュースとして拡散され，2 段階目では多くのユーザがそのニュースをフェイクだと認識することで，それ自体が別のニュースとして広まっていく過程からなるものとみなす．我々は，Twitter 上で拡散したフェイクニュースの 2 つのデータセットを作成し，本モデルの検証を行った．実験の結果，他の手法と比較して，本モデルがフェイクニュースの拡散過程を正確に予測できる点を示した．更に，テキスト分析によって，本モデルがユーザがそのニュースをフェイクだと気づき始める時間を適切に推定できることを示した．本モデルは，フェイクニュースの拡散ダイナミクスの理解に貢献するとともに，フェイクニュースの検出や緩和に役立つ可能性があると考えられる．

キーワード　Fake News, フェイクニュース，Point Process，点過程，拡散モデル，Twitter

1 はじめに

スマートフォンの普及とともに，人々は新聞やテレビなどの従来のメディアではなく，ソーシャルメディアからニュースを消費することが増加し，ともに，ソーシャルメディアによって様々な情報の共有や議論などが可能となった．一方で，ソーシャルメディアは社会に悪影響を及ぼす可能性のあるフェイクニュースの温床となっている．例えば，2016 年のアメリカ大統領選挙では，Twitter で投稿されたニュースの 25%がフェイクもしくは極端に偏っており，因果関係を分析すると，トランプ氏の支持者の活動がフェイクニュースの拡散に影響を与えていたと報告される [1]．選挙だけでなく，2011 年の東日本大震災などの自然災害 [2], [3] や，暴動や犯罪行為を引き起こすフェイクニュース [4] が世界各国で頻繁に共有されるようになった．

本研究では，Twitter でフェイクニュースはどのように拡散されるのかについての理解に取り組む．この問題は，社会科学上において重要であり，信頼できない情報や風評が社会にどのように拡散していくのかといった問いや，フェイクニュースの検出や緩和 [5], [6] にも応用可能であると考えられる．先行研究 [7], [8] では，フェイクニュースがソーシャルメディアで拡散する際の経路に着目され，拡散構造の側面が明らかにされてきた．しかしながら，フェイクニュースがオンライン上でどのように拡散していくのかといった時間的な側面についてはほとんど理解されていない．

我々は，これらの側面を理解するために，フェイクニュースが Twitter 上で 2 段階で拡散すると仮定する．1 段階目では，フェイクニュースは普通のニュースと拡散される．2 段階目では，多くのユーザが “correction time” の以後にそのニュースをフェイクと認識する，そして，そのニュース自体が別のニュースとして拡散される．本研究では，Twitter の拡散過程の予測やモデリングを行う Time-Dependent Hawkes process(TiDeH) [9] を拡張することで，上記の仮定を提案モデルとして定式化する．そして，提案モデルの有効性を確認するために，2 つの Twitter のフェイクニュースデータセットを作成し検証した．

本研究の貢献は以下の 3 つである．

- 本研究では，フェイクニュースの拡散が 2 段階でおこなわれるという仮定に基づいて簡易な点過程モデルを提案した．
- 提案モデルの有効性を検証するため，将来の投稿数の予測問題によって，提案モデルの予測性能について評価した．
- テキストマイニングの手法を用いて，提案モデルの有効性について確認した．

2 関連研究

オンライン上のコンテンツの人気や拡散の予測は数多く行われている [10], [11]．人気を予測する一般的な手法は機械学習を用いることで，分類問題 [12], [13] や回帰問題 [14] などに落とし込むことが多い．別の手法として数理モデルを開発し，訓練データセットを用いてモデルパラメータを適合させるものもあり，主に時系列モデルと点過程モデルの 2 つが存在する．時系列モデルは一定のウィンドウ幅における投稿の数について記

載するもので，ブログや Google Trends や Twitter での時間的活動を再現する SpikeM [15] やオンライン署名における広告の効果を考慮した時系列モデル [16] などが代表として挙げられる．点過程モデルは，情報拡散の self-exciting な性質を考慮して確率的に投稿時間を記述するものであり [17], [18]，拡散ダイナミックスにおけるネットワワーク構造やイベント時間の効果に関する理論的研究の動機づけともなっている [19]．最終的なシェアや投稿数の予測 [18], [20] やソーシャルメディアにおける時間的変動の予測 [9] のために，様々な点過程モデルが提案されている [18], [20]．さらに，これらのモデルは Youtube [21] や Twitter [22] での活動に対する外生的ショックを解釈するために適用される．提案モデルは，点過程モデルの 1 つである TiDeH [9] の拡張によって構築されており，我々の知る限り提案モデルはフェイクニュース拡散の新たな側面を取り込んだ初の定量的なモデルである．先行研究 [23] でもフェイクニュース拡散のモデルを提案しているが，これらは定性的な側面のモデル化に焦点を当てており，実際のデータセットを用いた予測性能の評価は行われていない，

本研究は，フェイクニュースの検出などにも関連する．フェイクニュースを自動的に検出する試みは数多く行われており，テキストの内容からフェイクニュースを検出するのが一般的である [5], [6]．Hassan ら [24] は文章から複数のカテゴリの特徴を抽出し，サポートベクトルマシンを適用してフェイクニュースを検出する．Rashkin ら [25] は long short-term memory (LSTM) をニュースのファクトチェックに活用する手法を提案している．提案モデルが活用するニュースの投稿や共有のタイミングなどのカスケードの時間情報によって，フェイクニュースや噂の検出や分類の性能を向上したという報告が数多くある [26] [27] [28] [29] [30] [31]．これらの研究は，提案モデルのようなフェイクニュース拡散のモデル化が検出や分類の性能を向上させる可能性があることを示唆する．また，深層学習モデルでは，時間情報の一部しか利用できない，もしくは，ユーザの反応の多いカスケードを扱えないという限界があるが，提案モデルのパラメータを時間表現のコンパクトな表現として利用することで，この限界を克服できる可能性がある．

3　フェイクニュース拡散のモデリング

フェイクニュース拡散について記述する点過程モデルを提案する．提案モデルの概要を図 1 に示す．提案モデルは以下の 2 つの仮定に基づく．

- ユーザは拡散の初期においてニュースをフェイクと認識しておらず，フェイクニュースは通常のニュースとして拡散される (図 1: 1st stage).
- ユーザは “correction time”t_c の周辺でニュースをフェイクだと認識することで，そのニュースがフェイクの性質を持って拡散される (図 1: 2nd stage).

要約すると，提案モデルはフェイクニュースの拡散が 2 つのカスケード (一連の情報やニュースによって引き起こされる投稿やシェア) によって構成されると仮定される：1) 通常のニュースとしてのカスケード，2) ニュースがフェイクであると認識されたことによるカスケード．これらのカスケードを記述するために，情報の減衰と概日リズムを考慮した TiDeH を用いる．

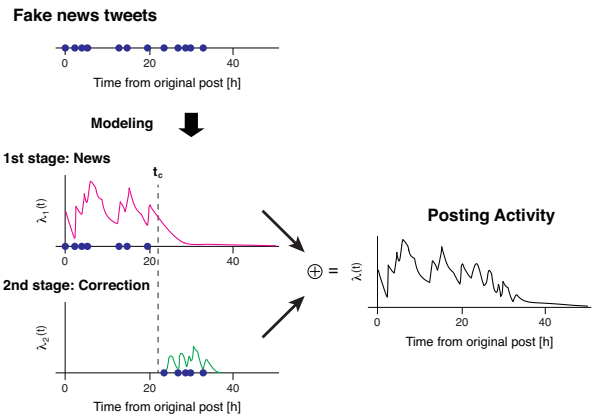

図 1: 提案モデルの概要図：フェイクニュースに関連した投稿がソーシャルメディア上でどのように拡散するのかを記述したモデルを提案する．青丸は投稿のタイムスタンプを示す．提案モデルはフェイクニュースを 2 段階で構成されると仮定する．最初は，フェイクニュースは通常のニュースとして拡散される (1st stage)．“correction time” である t_c を経て，Twitter ユーザはそのニュースがフェイクであると認識し，異なる性質を持ったニュースとして拡散する (2nd stage)．フェイクニュースの投稿カスケードは 2 段階の合計で算出される．

3.1　基礎モデル: Time-Dependent Hawkes process (TiDeH)

ニュースによる情報の拡散を記述する単一の点過程モデルについて述べる．点過程モデル [32] は小さな時間幅 $[t, t+\Delta t]$ における，投稿や共有の確率を $\lambda(t)\Delta t$ と記す．ここで，$\lambda(t)$ はカスケードの速度で，強度関数を意味する．TiDeH モデルの強度関数 [9] は以下のように以前の投稿に基づいて算出される:

$$\lambda_{\mathrm{TiDeH}}(t) = p(t)h(t), \tag{1}$$

メモリー関数 $h(t)$ は以下のように定義される:

$$h(t) = \sum_{i:t_i<t} d_i \phi(t - t_i), \tag{2}$$

$p(t)$ は infection rate，t_i は i 番目の投稿の時間，d_i は i 番目の投稿のフォロワー数を表す．Infection rate である $p(t)$ は、カスケードの情報の減衰と概日リズムという 2 つの特性を取り込んだもので，情報の感染力の強さを表す値である．TiDeH では $p(t)$ を以下のように算出する．

$$p(t) = a\left\{1 - r\sin\left(\frac{2\pi}{T_m}(t+\theta_0)\right)\right\}\mathrm{e}^{-(t-t_0)/\tau},$$

最初の投稿時間として $t_0 = 0$ として，振動の周期として $T_m = 24$ 時間として定義される．パラメータ a, r, θ_0, τ はそれぞれ，強さ，相対的な振動，振動のズレ，減衰項を意味する．メモリーカーネルである $\phi(t)$ は，フォロワーの反応時間の確率であり，ヘビーテイル分布になるように設定する [9], [18].

$$\phi(s) = \begin{cases} c_0 & (0 \leqq s \leqq s_0) \\ c_0(s/s_0)^{-(1+\gamma)} & (\text{Otherwise}) \end{cases}$$

各パラメータは以下のように設定する: $c_0 = 6.94 \times 10^{-4}$ (/seconds), $s_0 = 300$ seconds, $\gamma = 0.242$.

3.2 提案モデル

フェイクニュース拡散に関する点過程モデルを定式化する．ここで，フェイクニュースの拡散は，ニュースそのものに起因するものと，フェイクと明らかになったニュースに起因するものの 2 つのカスケードによって構成されると仮定する．フェイクニュースのカスケードは TiDeH を用いて 2 つのカスケードの合計として以下のように記述できる．

$$\lambda_{\mathrm{prop}}(t) = p_1(t)h_1(t) + p_2(t)h_2(t). \tag{3}$$

第一項目の $p_1(t)h_1(t)$ は，そのニュースによって引き起こされるカスケードの強度関数を示す．

$$\begin{aligned} p_1(t) =& a_1 \left\{1 + r \sin\left(\frac{2\pi}{T_m}(t+\theta_0)\right)\right\} \mathrm{e}^{-t/\tau_1}, \\ h_1(t) =& \sum_{i:t_i<\min(t,t_c)} d_i \phi(t-t_i), \end{aligned} \tag{4}$$

a_1 は，拡散におけるそのニュースの強さ，τ_1 は減衰項，$\min(t, t_c)$ はフェイクニュースが修正される時間である t_c と t の 2 つの値のうち小さい値を示す．第二項目の $p_2(t)h_2(t)$ は修正によって生じるカスケードであり，以下のように算出する．

$$\begin{aligned} p_2(t) =& a_2 \left\{1 + r \sin\left(\frac{2\pi}{T_m}(t+\theta_0)\right)\right\} \mathrm{e}^{-(t-t_c)/\tau_2}, \\ h_2(t) =& \sum_{i:t_c<t_i<t} d_i \phi(t-t_i), \end{aligned} \tag{5}$$

a_2 は，フェイクと明らかになった拡散におけるニュースの強さ，τ_2 は減衰項を表す．$p_2(t)$ の概日パラメータは $p_1(t)$ の概日パラメータと同様であると仮定する．数学的には，提案モデルの特殊なケース，具体的には以下の条件を満たすモデルが TiDeH である．

$$\tilde{a} = a_1 = a_2 e^{-t_c/\tilde{\tau}}, \quad \tilde{\tau} = \tau_1 = \tau_2 \tag{6}$$

式 (6) を式 (3)，(4)，(5) に代入することで、パラメータが $a = \tilde{a}$，$\tau = \tilde{\tau}$ となり提案モデルが TiDeH と同等であることが示される．

4 パラメータ推定

投稿の時系列からパラメータを推定する手法について述べる．提案モデルの 7 つのパラメータ $(a_1, \tau_1, a_2, \tau_2, r, \theta_0, t_c)$ は最尤推定法によって推定する．

$$l = \sum_i \log \lambda(t_i) - \int_0^{T_{\mathrm{obs}}} \lambda(s) ds, \tag{7}$$

t_i は i 番目の投稿時間，$\lambda(t)$ は式 (3) で計算される強さ，T_{obs} は観察時間を示す．我々は，はじめに correction time である t_c を $0.1T_{\mathrm{obs}} < t_c < 0.9T_{\mathrm{obs}}$ という制約の元，Brent 法 [34] で推定した．その後，推定された t_c を固定し，他のパラメータを $12 < \tau_1, \tau_2 < 2T_{\mathrm{obs}}$ (hours) という制約の元，を Scipy [1]による Newton 法 [33] を用いて推定した．

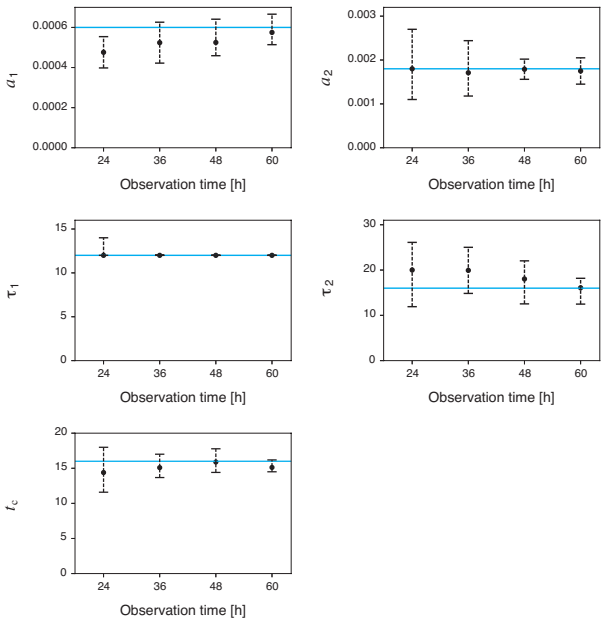

図 2: 観察時間に依存した，各パラメータ $(a_1, \tau_1, a_2, \tau_2, t_c)$ の推定精度の変化結果．100 回のシミュレーションによって推定された値を元に算出し，黒丸が中央値，エラーバーが四分位の範囲を示す．シアン色の線は真の値を表す ($a_1 = 0.0006$, $a_2 = 0.0018$, $\tau_1 = 12$, $\tau_2 = 16$, $t_c = 16$).

更に，パラメータ推定の妥当性を提案モデル式 (3) によって生成されたシミュレーションデータを用いて検証する．図 2 では，シミュレーションの結果として，観測時間がどのぐらい推定精度に影響を与えるのかを示す．パラメータの推定精度を評価するために，100 回のシミュレーションによる推定値の中央値と四分位の範囲を算出した．本シミュレーションによって，推定誤差は観測時間が長くなるにつれて減少することが明らかになった．この結果は，手法が 36 時間以上といった十分に長い観測時間で確実にパラメータを推定できることを示唆している．36 時間のシミュレーションデータから得られた絶対相対誤差の中央値は，a_1 で 18%，τ_1 で 11%，a_2 で 38%，τ_2 で 38%，t_c で 10%であった．1 つ目のカスケードパラメータ (a_1, τ_1) の推定精度と比較して，2 つ目のカスケードパラメータ (a_2, τ_2) の推定精度よりも悪い結果となった．これは 1 つ目のカスケードが全てのデータから推定されるのに対して，2 つ目のカスケードのパラメータは correction time(t_c) 以降のデータから推定されることによる，観測データの不足が原因だと考えられる．

加えて，$a_1 = a_2 e^{-t_c/\tau_2}$ かつ $\tau_1 = \tau_2$ の時，モデルのパラメータは一意に定まらない [35], [36]．提案モデルは $a_2 = 0$ かつ $t_c \geqq T_{\mathrm{obs}}$ のパラメータを持つ TiDeH と同等になり，他のパラメータでも観測データの再現が可能になるからである．図 3 に $a_1 = a_2 e^{-t_c/\tau_2}$ かつ $\tau_1 = \tau_2$ のシミュレーション結果を示す．

5 データセット

フェイクニュースに関する 2 つのデータセットを用いて，提

1 : https://docs.scipy.org

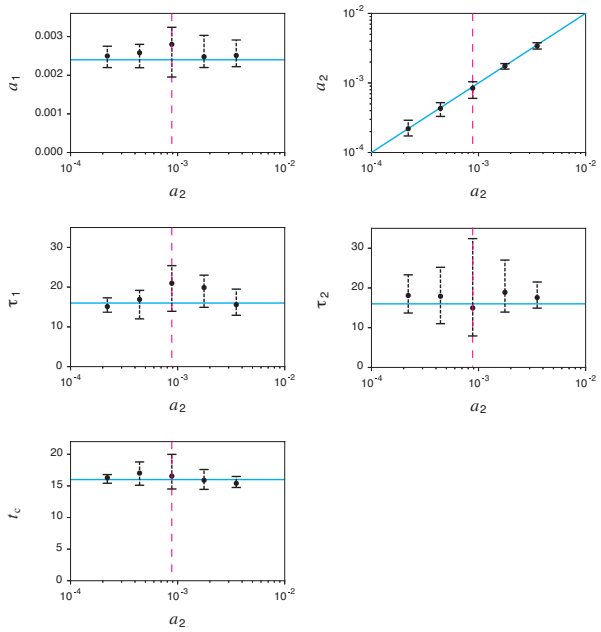

図 3: $a_2 = a_1 e^{-t_c/\tau_2}$ かつ $\tau_1 = \tau_2$ というパラメータが特定できない時のパラメータ推定精度．100 回のシミュレーションによって推定された値を元に算出し，黒丸が中央値，エラーバーが四分位の範囲を示す．マゼンダの点線はパラメータが特定できない設定を示す．シアンの直線は真の値を表す ($a_1 = 0.0024$，$\tau_1 = \tau_2 = 16$，$t_c = 16$，a_2 は 2.2×10^{-4} から 3.5×10^{-3} の間を変化させる．)

案モデルの評価と有用性を検討する．Twitter におけるフェイクニュースのデータセットはいくつか公開されているが [37], [38]，これらのデータセットはリツイートなどが追えているわけではなく，完全なデータセットではないことが多い．このことからも，詳細に情報の拡散を追うために，Twitter から手動で Recent Fake News(RFN) と東日本大震災データセット (Tohoku) の 2 つのデータセットを作成した．作成したデータセットのうち，RFN では 61%が Tohoku では 20%が最初の投稿のリツイートとなっている．

5.1 Recent Fake News データセット (RFN)

2019 年の 3 月から 5 月に，米国のフェイクニュース検証である Politifact.com [2]と Snopes.com [3]の 2 つのサイトで報告された 10 のフェイクニュースについて収集した．PolitiFact は主に米国の政治ニュースや政治家の発言を対象とした独立で無党派のオンラインファクトチェッキングサイトである．Snopes は初期の検証サイトの 1 つで，政治や社会的イベントや話題となっている事柄を取り扱うサイトである．Twitter API を用いて，URL とキーワードを元にフェイクニュースとのその関連の高い投稿を収集した．更に，収集したニュースから以下の 2 つの条件に基づいて 6 つのフェイクニュースを選択した．1) 投稿数が 300 以上，2) シミュレーションによるデータを用いた実験から，観測期間が 36 時間以上であるもの (4 章)．

5.2 東日本大震災データセット (Tohoku)

2011 年 3 月 11 日に発生した東日本大震災では，多くのフェイクニュースが発生した．Twitter API を用いて 2011 年 3 月 12 日から 3 月 24 日の期間に収集した日本語ツイートから，フェイクニュース検証記事 [4]に基づいてデータセットを作成した．RFN データセットと同様の条件で，19 件のフェイクニュースを抽出した．収集したニュースの概要は github [5]に掲載する．

6 実験と結果

提案モデルを評価するために，以下の Twitter の投稿量の予測タスクに取り組む．フェイクニュースの拡散として，最初の投稿 ($t_0 = 0$) から時間 $T_{\rm obs}$ までの投稿列 $\{t_i, d_i\}$ を実験データとして用いる．ここでは，i 番目の投稿した時間を t_i，i 番目投稿したユーザのフォロワー数を d_i，$T_{\rm obs}$ は観察時間を示す．実験では，t_0 から $T_{\rm obs}$ までの期間をパラメータ推定時間，$T_{\rm obs}$ から $T_{\rm max}$ までの期間をテスト期間と設定し，対象ニュースの 1 時間ごとの投稿数を予測するタスクに取り組む．本章では，実験設定と予測方法について説明し，予測実験によってベースラインと提案手法の性能を比較する．

6.1 実験設定

そのニュースが含まれた投稿の全期間 $[0, T_{\rm max}]$ を訓練とテスト期間に分割する．訓練期間はその全期間の前半部分 $[0, 0.5T_{\rm max}]$，後半部分 $[0.5T_{\rm max}, T_{\rm max}]$ をテスト期間として設定する．予測性能は，平均絶対誤差 (Mean absolute error) と中央絶対誤差 (Median absolute error) によって評価する．これらの評価指標は以下の計算式で示される．

$$\text{Mean Absolute Error} = \frac{1}{n_b}\sum_k |\hat{N}_k - N_k|,$$

$$\text{Median Absolute Error} = \text{Median}(|\hat{N}_k - N_k|) \quad (k = 1, 2, \cdots n_b),$$

ここでは，$\hat{N}_k$ が k 番目 のビン $[(k-1)\Delta + T_{\rm obs}, k\Delta + T_{\rm obs}]$ における予測値，N_k が同じビンにおける真の値を指し示す．n_b はビンの数を，ビン幅である $\Delta = 1$ 時間として設定する．

6.2 提案モデルによる予測方法

まず，訓練期間のデータから最尤法を用いてモデルパラメータを推定する (4 章参照)．次に，テスト期間 ($t \in [T_{\rm obs}, T_{\rm max}]$) の強度関数 $\hat{\lambda}(t)$ を算出する．

$$\hat{\lambda}_{\rm prop}(t) = \hat{\lambda}_1(t) + \hat{\lambda}_2(t) \tag{8}$$

$$\hat{\lambda}_1(t) = p_1(t) \sum_{i:t_i<t_c} d_i \phi(t - t_i), \tag{9}$$

ここでは，$\hat{\lambda}_1(t)$ は $\hat{\lambda}_2(t)$ それぞれ，1 つ目と 2 つ目それぞれのカスケードの強度関数を表す．1 つ目のカスケードの強度関数 $\hat{\lambda}_1(t)$ は，推定されたパラメータ ($\{a_1, \tau_1, r, \theta_0\}$) と推定された corretion time (t_c) 以前の観察データ $\{t_i, d_i\}$ に基づいて計算

2 : https://www.politifact.com/
3 : https://www.snopes.com/
4 : https://blogos.com/article/2530/
5 : https://github.com/hkefka385/extended_tideh

される．Tohoku データセットでは，フォロワーの人数の取得ができなかったため，d_i を 1 に固定する．2 つ目のカスケードの強度関数 $\hat{\lambda}_2(t)$ は以下の積分によって算出できる．

$$\hat{\lambda}_2(t) = f(t) + d_p p_2(t) \int_{T_{\text{obs}}}^{t} \hat{\lambda}_2(s)\phi(t-s)ds, \qquad (10)$$

$$f(t) = p_2(t) \sum_{i:t_c<t_i<T_{\text{obs}}} d_i \phi(t-t_i),$$

d_p は訓練期間中の平均フォロワー人数とする．

6.3 予測結果

提案モデルの予測性能を，線形回帰 (LR) [14]，強化ポアソン過程 (RPP) [40]，TiDeH [9] の 3 つのベースラインと比較する．図 4 に，フェイクニュースの累積投稿数の時系列とその予測結果の 3 つの結果を示す．提案モデル (図 4 のマゼンダ) は他のベースラインよりも実際の時系列に近い結果となることが示された．また，提案モデルは投稿活動の減衰効果を再現している一方で，ベースラインモデルは投稿数を過大評価する傾向が見られた．

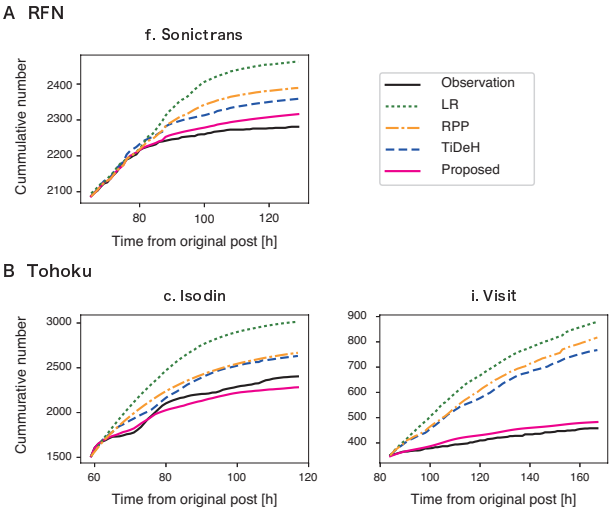

図 4: (A)RFN，(B)Tohoku データセットの一部の例における予測結果，累積数の時系列を示す．緑，オレンジ，青の破線はそれぞれベースライン (LR，RPP，TiDeH) の予測結果を示す．黒線とマゼンダ線は，実際の累積数と提案モデルの予測結果を示す．

次に，提案モデルのパラメータの分布について検証する．ニュースの拡散の強さを示す a の分布について，2 段階目のカスケードにあたる誤りと明らかになったニュースの拡散は，1 段階目にあたる通常のニュースとしての拡散よりも弱い傾向があった．具体的には RFN データセットの 67%，Tohoku データセットの 79%がそのような傾向が見られる．これは，ニュースがフェイクと明らかになった事実よりも，ニュースの内容自体がユーザにとて意外性が大きいことに起因すると考えられる．次に，減衰を表す τ について検証する．1 段階目カスケードの τ_1 は両方のデータセットで約 40 時間程度であった．具体的には RFN データセットの中央値が約 35 時間 (22–92 時間)，Tohoku データセットでの中央値が約 40 時間 (19–54 時間) であった．2 段階目のカスケードの τ_2 は，両データセットとともに広く分布しており，シミュレーションデータで観測された結果と一致している (図 2)．“correction time” である t_c は最初の投稿から 30 時間から 40 時間程度になる傾向がある．RFN データセットでは中央値 32 時間 (21–54 時間)，Tohoku データセットでは 37 時間 (31–61 時間) であった．先行研究 [41] では，フェイクニュース検証サイトがフェイクニュースを検知するのは，元の投稿から約 10–20 時間後であることが報告されている．この報告から，Twitter ユーザはフェイクニュース検証サイトの報告から約 10–20 時間後にフェイクニュースを認識していることが示唆される．

最後に，2 つのフェイクニュースデータセットの予測性能を評価した結果を表 1 に示す．2 つのデータセットと評価指標で，提案手法はベースラインよりも高い精度となった．具体的には，提案モデルが TiDeH と比較して，RFN と Tohoku データセットにおいてそれぞれ 32%，42%誤差 (Mean) が小さくなっている．また，これまでの研究 [9], [18], と同様に，点過程モデルに基づく手法 (提案手法，TiDeH，RPP) が線形回帰法 (LR) よりも優れた結果を示した．実際，提案モデルはほとんどのフェイクニュース (RFN の 100%，Tohoku の 89%) で最も高い精度を達成した．しかしながら，TiDeH は一部データで提案モデルよりも精度が高いという結果になった．これは，一部のフェイクニュースが 2 つのカスケードでなく 1 つのカスケードで拡散することがあるためだと考えられる．例えば，エイプリールのようにユーザがそのニュースをフェイクだと事前に知っている場合や，ユーザがそのニュースの間違いに興味が無い場合に生じる可能性がある．これらの結果から，提案手法は Twitter 上でのフェイクニュース関連投稿の拡散予測に有効であることが示された．

表 1: 2 つのデータセットでの予測性能: 1 時間あたりの絶対誤差の平均値 (Mean) と中央値 (Median)．最高精度を太字で示す．

Datasets	RFN		Tohoku	
Metric	Mean	Median	Mean	Median
LR	88.3	5.08	13.9	4.51
RPP	61.8	3.12	8.23	2.30
TiDeH	54.2	1.89	4.12	1.99
Proposed	**36.9**	**1.37**	**2.40**	**1.80**

7 correction time の推定

本研究では，提案手法が既存手法よりもフェイクニュース拡散予測が可能であることを示した．提案モデルでは，Twitter ユーザが correction time である t_c に拡散されているニュースがフェイクであると気づくと仮定している．本節では，テキストマイニングを通してこの仮説を検証する．

はじめに，フェイクニュース指摘に関連する言葉の頻度

(“FakeFrequency” と呼称) と時間 t_c の比較を図 5 を示す．“FakeFrequency” とは，1 時間あたりに投稿された “fake”，“false”，“not true”，“not real” といったフェイクに関する語を含む投稿件数を示す．RFN データセットの “FakeFrequency” は Tohoku のものよりも少ない傾向にあり，具体的には，RFN データセットの b. Notredome には 29 件，f. Socnitrans には 277 件，Tohoku データセットの a. Saveenergy に 1752 件，l. Taiwan に 1616 件，q. Cartoonist に 1723 件，s. Turkey に 1930 件の “FakeFrequency” がみられた．これは，RFN データセットの投稿のほとんどが元の投稿のリツイートによることが原因だと考えられる．観察の結果，“correction time” 周辺で “FakeFrequency” が増加していることを確認できる．特に，図 5 が示すように Tohoku データセットの l. Taiwan や q. Cartoonist で，“correction time” と同時期にフェイクニュースに関する語が出現する．

次に，図 6 に，Tohoku データセットの s. Turkey で correction time である t_c の前後で，投稿の内容がどのように変化したのかを WordCloud で可視化しものを示す．このフェイクニュースは，トルコが日本に 100 億円の支援を行ったという内容である．correction time 前では，“pro-japanese” という語が示すようにトルコが親日国であるという情報とともに広まったことが示唆される．correction time 以後には，新たに “False rumor” という語や “Taiwan” という語が出現しており，これは「台湾」に関する別のフェイクニュースと関連して出現したことを示している．これらの結果は，Twitter ユーザが correction time 以後にニュースがフェイクであると認識したことを示唆しており，提案モデルの仮説を支持するものである．

8 おわりに

本研究では，Twitter におけるフェイクニュースに関する将来の投稿量を予測可能な点過程モデルを提案した．提案モデルを構築するために，フェイクニュースの拡散は 2 段階で構成されると仮定した．最初は，フェイクニュースは通常のニュースとして拡散される (1st stage)，次に correction time である t_c を経て，Twitter ユーザはそのニュースがフェイクであると認識し，異なる性質を持ったニュースとして拡散する (2nd stage) というものである．実験では，フェイクニュースのデータセットを 2 つ作成し，フェイクニュースの将来の拡散を予測した結果，他の既存手法よりも提案モデルが正確に予測できることを示した．さらに，テキストマイニングを用いて，提案モデルによって推定される correction time がユーザがニュースをフェイクと認識する時間であることを示した．

将来の展望として，2 つの方向性があると考える．1 つ目として，本研究ではフェイクニュースの拡散を 2 段階であると仮定しているが，更に詳細にモデリングすることも可能である．例えば，複数の投稿や隠れ変数を考慮することで，1st stage から 2nd stage への自然な切り替えを組み込むことができる．2 つ目として，提案モデルをフェイクニュースの検出や緩和といった実用的な問題に適用できる．提案モデルはフェイクニュース

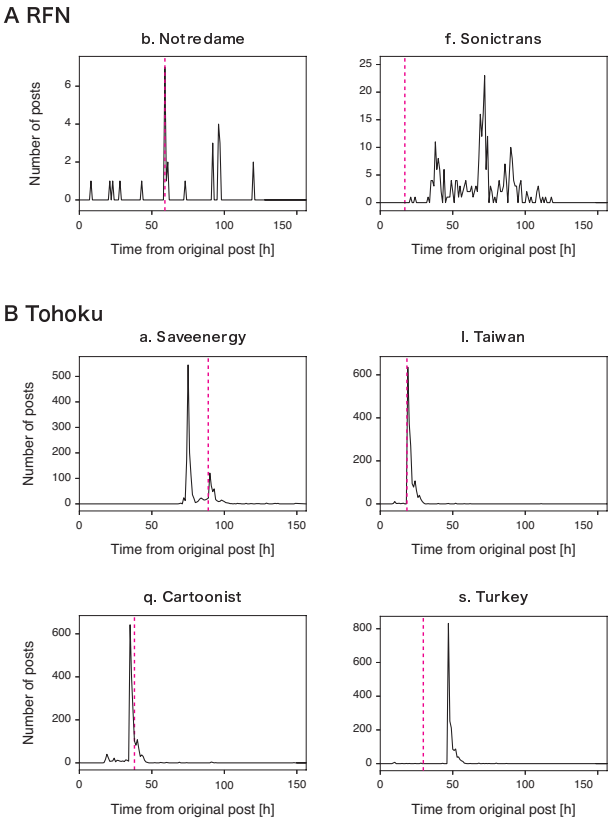

図 5: フェイクニュースであることを指摘すると考えられる語の出現回数 “FakeFrequency” を示した時系列．(A) RFN，(B) Tohoku データセット．黒線は 1 時間あたりの “FakeFrequency” の時系列，マゼンタの縦線は correction time である t_c を示す．

図 6: Tohoku データセット s. Turkey における，correction time である t_c 以前 (左) と以後 (右) の投稿によって作成した WordCloud．それぞれ，投稿でよく用いられた単語上位 10 語を示す．

拡散のモデル化を実現するとともに，フェイクニュースの拡散に関連する時間情報をコンパクトに表現するという点で有益である．

謝　辞

本研究の一部は，JSPS 科研費 JP19K20279, JP19H04221, JP17H03279, JP18K11560, JP19H01133，厚生労働省科学研究費補助金（課題番号：H30-新興行政-指定-004），および JST，さきがけ，JPMJPR1925 の支援を受けたものです．

文　献

[1] Bovet A, Makse HA. (2019) Influence of fake news in Twitter during the 2016 US presidential election. Nature communications 10: 1-14.
[2] Takayasu M, Sato K, Sano Y, Yamada K, Miura W,

Takayasu H. (2015) Rumor diffusion and convergence during the 3.11 earthquake: a Twitter case study. PLoS one 10: e0121443.

[3] Hashimoto T, Shepard DL, Kuboyama T, Shin K, Kobayashi R, Uno T. (2020) Analyzing temporal patterns of topic diversity using graph clustering. The Journal of Supercomputing. https://doi.org/10.1007/s11227-020-03433-5.

[4] Marc F, Cox JW, Hermann P. (2016) Pizzagate: From rumor, to hashtag, to gunfire in dc. Washington Post.

[5] Shu K, Sliva A, Wang S, Tang J, Liu H. (2017) Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter 19: 22-36.

[6] Sharma K, Qian F, Jiang H, Ruchansky N, Zhang,M, Liu Y. (2019) Combating fake news: A survey on identification and mitigation techniques. ACM Transactions on Intelligent Systems and Technology 10: 1-42.

[7] Vosoughi S, Roy D, Aral S. (2018) The spread of true and false news online. Science 359: 1146-51.

[8] Zhao Z, Zhao J, Sano Y, Levy O, Takayasu H, Takayasu M, Li D, Wu J, Havlin S. (2020) Fake news propagates differently from real news even at early stages of spreading. EPJ Data Science 9:7.

[9] Kobayashi R, Lambiotte R. (2016) TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. Proceedings of 10th International Conference on Web and Social Media, ICWSM 2016. p. 191-200.

[10] Kursuncu U, Gaur M, Lokala U, Thirunarayan K, Sheth A, Arpinar IB. (2019) Predictive analysis on Twitter: Techniques and applications. Emerging research challenges and opportunities in computational social network analysis and mining, p. 67-104. Springer, Cham.

[11] Tatar A, De Amorim MD, Fdida S, Antoniadis P. (2014) A survey on predicting the popularity of web content. Journal of Internet Services and Applications 5:8.

[12] Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J. (2014) Can cascades be predicted? Proceedings of the 23rd international conference on world wide web, WWW 2014, p. 925-936.

[13] Petrovic S, Osborne M, Lavrenko V. (2011) Rt to win! predicting message propagation in twitter. International Conference on Web and Social Media, ICWSM 2011, p.586-589.

[14] Szabo G, Huberman BA. (2010). Predicting the popularity of online content. Communications of the ACM, 53.8: 80-88.

[15] Matsubara Y, Sakurai Y, Prakash BA, Li L, Faloutsos C. (2012) Rise and fall patterns of information diffusion: model and implications. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD 2012, p.6-14

[16] Proskurnia J, Grabowicz P, Kobayashi R, Castillo C, Cudré-Mauroux P, Aberer K. (2017) Predicting the success of online petitions leveraging multidimensional time-series. Proceedings of the 26th International Conference on World Wide Web, WWW 2017, p. 755-764.

[17] Masuda N, Takaguchi T, Sato N, Yano K. (2013) Self-exciting point process modeling of conversation event sequences. Temporal networks, pp. 245-264. Springer, Berlin, Heidelberg.

[18] Zhao Q, Erdogdu MA, He HY, Rajaraman A, Leskovec J. (2015) Seismic: A self-exciting point process model for predicting tweet popularity. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2015, p.1513-1522.

[19] Delvenne JC, Lambiotte R, Rocha LE. (2015) Diffusion on networked systems is a question of time or structure. Nature communications. 6.1:1-10.

[20] Medvedev AN, Delvenne JC, Lambiotte R. (2019) Modelling structure and predicting dynamics of discussion threads in online boards. Journal of Complex Networks.7.1:67-82.

[21] Rizoiu MA, Xie L, Sanner S, Cebrian M, Yu H, Van Hentenryck P. (2017) Expecting to be HIP: Hawkes intensity processes for social media popularity. Proceedings of the 26th International Conference on World Wide Web, WWW 2017 p.735-744.

[22] Fujita K, Medvedev A, Koyama S, Lambiotte R, Shinomoto S. (2018) Identifying exogenous and endogenous activity in social media. Physical Review E. 98.5:052304.

[23] Törnberg P. (2018) Echo chambers and viral misinformation: Modeling fake news as complex contagion. PLoS one. 13.9:e0203958.

[24] Hassan N, Arslan F, Li C, Tremayne M. (2017) 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, KDD 2017, p.1803-1812.

[25] Rashkin H, Choi E, Jang JY, Volkova S, Choi Y. (2017) Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the 2017 conference on Empirical Methods in Natural Language Processing, EMNLP 2017, p.2931-2937.

[26] Kwon S, Cha M, Jung K, Chen W, Wang Y. (2013) Prominent features of rumor propagation in online social media. Proceedings of the 2013 IEEE 13th International Conference on Data Mining, ICDM 2013, p.1103-1108.

[27] Ruchansky N, Seo S, Liu Y. (2017) Csi: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, p.797-806.

[28] Lukasik M, Srijith PK, Vu D, Bontcheva K, Zubiaga A, Cohn T. (2016) Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, p.393-398.

[29] Farajtabar M, Rodriguez MG, Zamani M, Du N, Zha H, Song L. (2015) Back to the past: Source identification in diffusion networks from partially observed cascades. Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, AISTATS 2015, p.232-240.

[30] Dutta HS, Dutta VR, Adhikary A, Chakraborty T. (2020) HawkesEye: Detecting fake retweeters using Hawkes process and topic modeling. IEEE Transactions on Information Forensics and Security. 15: p. 2667-2678.

[31] Murayama T, Wakamiya S, Aramaki E. (2020) Fake News Detection using Temporal Features Extracted via Point Process. Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media.

[32] Daley DJ, Vere-Jones D. (2003) An introduction to the theory of point processes, volume 1: Elementary theory and methods. Verlag New York Berlin Heidelberg: Springer.

[33] Nash SG. (1984) Newton-type minimization via the Lanczos method. SIAM Journal on Numerical Analysis 21: p. 770-788.

[34] Brent RP. (2013) Algorithms for minimization without derivatives. Dover Publications.

[35] Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics 25: p. 1923-1929.

[36] Gontier C, Pfister JP. (2020) Identifiability of a Binomial Synapse. Frontiers in computational neuroscience 14:86.

[37] Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. (2020) FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. Big Data 8: p. 171-188.

[38] Ma J, Gao W, Mitra P, Kwon S, Jansen BJ, Wong KF, Cha M. (2016). Detecting rumors from microblogs with recurrent neural networks. Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016, p. 3818-3824.

[39] Usui S, Toriumi F, Hirayama T, Enokibori Y, Mase K. (2015) Why did false rumors diffuse after the 2011 earthquake off the pacific coast of Tohoku? Impact analysis of the network structure. Electronics and Communications in Japan 98: p. 1-13.

[40] Gao S, Ma J, Chen Z. (2015) Modeling and predicting retweeting dynamics on microblogging platforms. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, p. 107-116.

[41] Shao C, Ciampaglia GL, Flammini A, Menczer F. (2016) Hoaxy: A platform for tracking online misinformation. Proceedings of the 25th international conference companion on world wide web, WWW 2016, p. 745-750.