様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ８ 日現在

機関番号：１２６１２
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１８８３

研究課題名（和文）情報共有モデルに基づくフェイクニュースの検知、予測および対策に関する研究

研究課題名（英文）Detecting, predicting, and deterring fake news using information-sharing models

研究代表者
吉浦　裕（Yoshiura, Hiroshi）

電気通信大学・その他部局等・名誉教授

研究者番号：４０３６１８２８

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：フェイクニュースが大きな社会問題になっているが、従来の対策には、検知の迂回が可能という技術的問題に加え、正しい検知結果をユーザが受け入れないという心理的問題があった。また、フェイクニュースの拡散予測の研究が欠落していた。そこで、ネットワークにおけるユーザ間の情報共有を表現するモデルに基づいて、対策を検討した。提案法は、実ネットワークにおける検査対象ニュースの伝搬をモデル上で再現し、その時のモデルのパラメータ値に基づいてニュースの信頼度を推定すると共に、フェイクニュースを受け入れたユーザの心理状態を推定し、対策することが可能であり、フェイクニュースの今後の広がりを予測することができる。

研究成果の学術的意義や社会的意義
情報共有モデルを用いたフェイクニュース検知は従来研究されていない。提案法は、ニュースとユーザの全体的な関係に基づくため回避が困難であり、モデルに手続き的なプログラムを組み込み可能であるため拡張性が高い。フェイクニュース対策の最も困難な課題は、ユーザが真実を受け入れないことであるが、提案法は、モデルのパラメータ値からネットワークユーザの心理状態を推定し、この問題への対策の基礎情報を得ることができる。また、従来研究されなかったフェイクニュースの拡散予測をモデル上で初めて可能にした。フェイクニュースの拡散は極めて深刻な社会問題であり、本研究の社会的意義は大きい。

研究成果の概要（英文）：Fake news cause serious problems in modern society. Conventional countermeasures could be circumvented by knowledgeable fake news writers. In addition, network users did not accept even correct detection of fake news due to the backfire effect. We proposed countermeasures based on models representing how information disseminates among network users. Our proposed methods conform news dissemination simulated on the models to real dissemination data by optimizing the model parameters. The proposed methods estimate reliability of news as well as belief of network users based on the optimized parameters, and estimate future dissemination of fake news by continuing the simulation.

研究分野：情報セキュリティ

キーワード：フェイクニュース　誤情報　検知　拡散予測　社会分断

様　式　C－１９、F－１９－１（共通）

１．研究開始当初の背景

テレビや新聞等のマスメディアの利用が減少し、ソーシャルメディアを情報源とする人が増えるなか、フェイクニュースが大きな社会問題になりつつあった。フェイクニュースおよびそれを受け入れる人々の特徴を社会学および心理学の面から調査分析する研究によると、フェイクニュースを信じる人は「事実を指摘しても受け入れない」傾向があり、対策は容易ではないことが明らかになってきていた。

当時のフェイクニュース対策のほとんどは、フェイクニュースの検知およびフェイクニュースの拡散形態の分析であった。フェイクニュースの検知は、あるニュースがフェイクかリアルかの判別であり、２種類の手法が研究されていた。第１の手法はニュースのコンテンツ（文章、画像等）や伝搬特性（広がる速さ等）に基づくが、フェイクニュースの発信者が検知手法を推定し迂回できるという問題があった。第２の手法は、複数のニュースとそれらを発信または転送した複数の人を観測事象とし、人とニュースの信用度をパラメータとし、観測事象の尤度を最大化するようにパラメータを最適化することで、ニュースの信頼度を推定していた。この手法は、人とニュースの全体的な関係（信用できない人は信用できないニュースにかかわる可能性が高い）を利用するので迂回は困難であるが、尤度を数式で表現する必要があるので柔軟性に乏しかった。たとえば、他の手法との融合や検知者の知識（ニュース発信者の意図など）の利用が困難であった。さらに、上記の手法のいずれも、「事実を指摘しても受け入れない」という問題に対処できないため、実効性に限界があることが明らかになりつつあった。

フェイクニュースの拡散形態の研究では、ニュースの内容とその拡散の広さ・速度の関係が分析されている。しかし、出現したフェイクニュースがこれからどのように拡散するかを推定する研究は行われていなかったので実用性が乏しかった。また、フェイクニュースに限らず、ネットワーク上の多人数に対する攻撃は、最初は単純な動機から始まるが、次第に高度な計画性を有する形態に変わってくる。たとえばマルウェアを用いる攻撃も、自己顕示から金銭目的、国家レベルの攻撃へと変わってきた。フェイクニュースの社会に与える影響の大きさを考慮すると、今後の高度な攻撃とその効果を予想し、警鐘を鳴らす必要があるが、当時は、その種の研究は行われていなかった。

２．研究の目的

研究開始当初の研究目的は以下であった。

(1) 迂回が困難で、他の手法との融合や検知者の知識の利用が容易なフェイクニュースの検知手法の確立
(2)「事実を指摘しても受け入れない」問題に対応する手法の探求
(3) フェイクニュースの拡散予測手法の探求

３．研究の方法

マルチエージェントシステムの分野では、ユーザをノード、ユーザ間の関係（Twitter/X のフォロワー関係等）をリンクとしたネットワーク構造において、ユーザ間の情報伝搬を表現・分析する情報共有モデルが長く研究されている。たとえば、図１のように、Pryymak らは、ニュースの真偽、ニュース提供者の信頼度、ユーザ間の信頼度、各ユーザのニュースを信じる度合をパラメータとして、ニュースの共有をモデル化している[1]。これらの情報共有モデルは、意図的なフェイクニュースは考慮していないが、誤ったニュースの影響をモデル化しており、フェイクニュースの伝搬を表現・分析するための拡張が可能である。

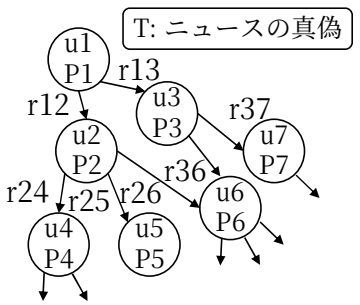

図１　情報共有モデルの概要

そこで、上記の研究目的１（迂回困難で柔軟な検知）、２（事実の指摘を受け入れない問題の解決）、３（拡散予測）について、情報共有モデルに基づく検討を行った。実ネットワークにおける検査対象ニュースの伝搬をモデル上で再現できれば、その時点のモデルのパラメータ値は検査対象ニュースの真偽、ニュース伝搬時のユーザ間の信頼度、ユーザのニュースを信じる度合等に対応している。そこで、モデルのパラメータ値に基づいて、フェイクニュースを検知できる。この検知手法は、人とニュースの全体的な関係に基づくので回避が困難である。また、ニュースの伝搬や信頼度の更新の規則をプログラムによって記述することができるので、他の手法との融合や検知者の知識の利用が容易であり、研究目的１の達成に適すると考えた。

また、検知時点のモデルのパラメータ値を用いて、ユーザ毎のフェイクニュースへの信じ込みの度合やユーザ間の信頼度を予測可能となり、「事実を受け入れない」問題への対処の基礎情報を得ることができるので、研究目的２の達成に資する。さらに、インフルエンサーのアカウントの乗っ取りやなりすまし、信頼度の偽装などを情報共有モデル上で表現し、フェイクニュースの

拡散をシミュレーションすることで、研究目的３の達成に資する。

本研究課題の申請後、フェイクニュースはテキストによるものだけでなく、画像を用いるものが急増した。また、大規模言語モデル（LLM）を用いた自然言語処理および生成系 AI の出現と急激な発展があった。そこで、画像を用いたフェイクニュースへの対策、および、LLM／生成系 AI を用いて作成されたフェイクニュースへの対策を研究目的に加えた。研究代表者らは、ソーシャルメディアの匿名投稿文から個人を特定する技術を長年研究している。一方、近年、深層学習に基づいて画像のキーワードを抽出する技術が Web API として利用可能となっている。この画像解析 API を投稿文からの個人特定技術に組み込むことで、投稿画像からの投稿者特定が可能になると考えた。また、生成系 AI によるフェイクニュース作成への対策として、生成系 AI の作成したソーシャルメディアの投稿と人手で作成した投稿を識別する手法を検討した。この識別において、公開されている LLM を活用することにした。

研究目的１および２の達成評価用のデータとして、フェイクニュース 11 件、リアルニュース 10 件、それらのニュースを発信・転送した Twitter ユーザ 10,438 人のアカウント、アカウント間のフォローフォロワー関係を収集した。また、LLM／生成系 AI の作成したフェイクニュースへの対策を評価するために GPT-4 の作成した 4,207 件の Twitter ニュース、人手で作成した 12,714 件の Twitter ニュースを収集した。なお、研究目的３の達成評価は Stanford 大学が公開している Facebook および Twitter のデータ、投稿画像からの投稿者特定の評価には、研究代表者らの従来研究で用いていた 51 件の Twitter アカウントとその投稿文を用いた。

４．研究成果

(1) 迂回困難で柔軟な拡張が可能なフェイクニュースの検知手法の確立

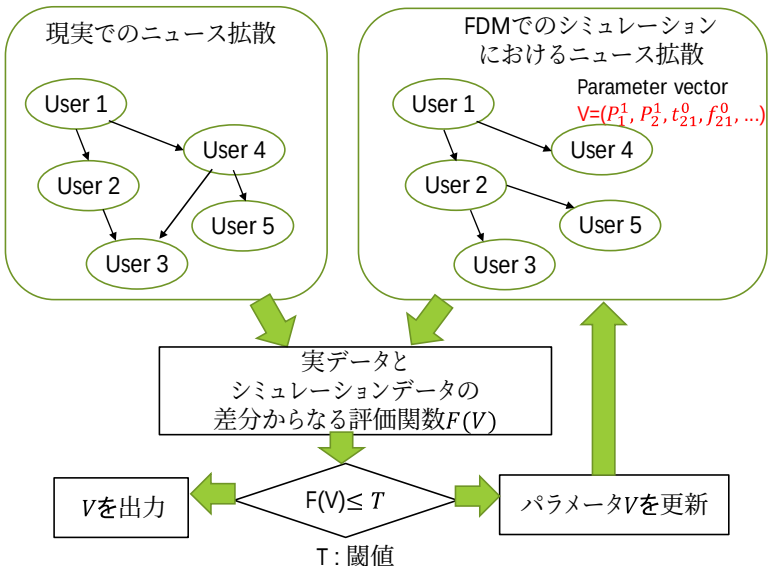

図２ 情報共有モデルを用いたフェイクニュース検知方式

Pryymak らの情報共有モデルを拡張して、フェイクニュースを扱うための情報共有モデル FDM（Fake news Dissemination Model）を開発した。Pryymak らのモデルでは、ニュースの誤りは意図的ではなくミスによるとしており、ニュースの真偽は確率的に設定していた。本研究では、フェイクニュースの発信者は意図的に嘘のニュースを流すとし、フェイクニュース発信者は確率 1.0 で誤ったニュースを流し、リアルニュースの発信者は確率 1.0 で正しいニュースを流すモデルとした。また、Ground Truth（ニュースが実際に真か偽か）をモデル内のニュース発信者は知っており、一般ユーザは知らないとした。ユーザが真実を信じやすい傾向および嘘を信じやすい傾向を想定して、Ground Truth が真か偽かに依存して、各ユーザのニュースを信じる度合いの初期値を確率的に設定した。

この FDM 上でシミュレーションしたニュースの拡散と現実のニュース拡散が最も近くなるように、FDM のパラメータであるニュースの真偽、ニュース提供者の信頼度、ユーザ間の信頼度と不信度、各ユーザのニュースを信じる度合を最適化し、収束時のパラメータ値からニュースの真偽を判定する手法を提案した（図２）。

フェイクニュースとリアルニュース各１件を用い、Ground Truth=(Fake, Real)として、予備実験を行った結果、(R,R)、(R,F)、(F,R)、(F,F)の４通りの予測のうち(R,F)、(F,R)の評価値 F(V)が(R,R)、(F,F)の評価値よりも小さかったことから、2 つのニュースの一方が Real で他方が Fake であることを検知できた。

提案法は、マルチエージェント分野で長年研究されてきた情報共有モデルをフェイクニュース検知に応用した初めての試みである。また、人とニュースの全体的な関係を利用するので回避が困難で、手続きの組み込みによる柔軟な拡張が可能である。提案法は(R,F)と(F,R)に対して同じ評価値を出力するので、2 つのニュースの真偽が異なることは検知できるが、どちらが真であるかの判断には別手法の兼用が必要である。また、計算量がニュース数の指数に比例する。これらの点について今後改良したい。

(2) 「事実を指摘しても受け入れない」問題への対応

Ground Truth である(F,R)を予測した時の、各ユーザがニュースを信じる度合の分布を図３に示す。図の横軸はニュースを信じる度合、縦軸はその度合いを有するユーザの人数を表す。この図より、ユーザはフェイクニュースに対して真実のニュースよりも極端な意見を持ちやすいこ

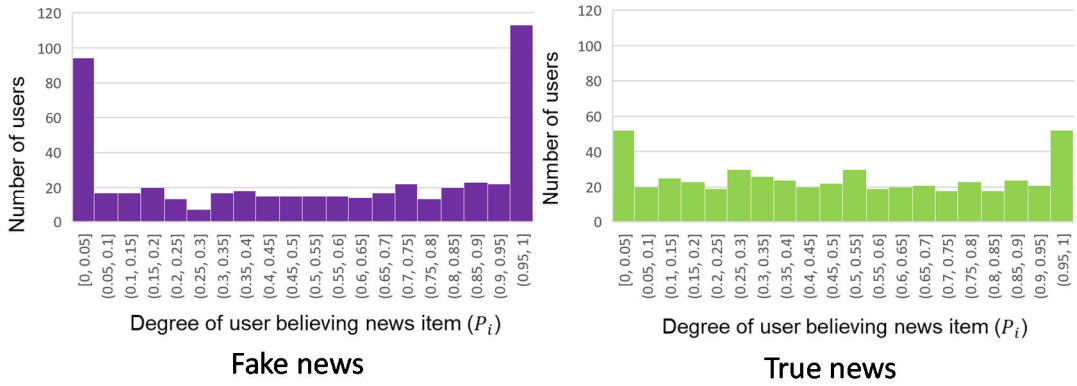

図３　ニュースを信じる度合のユーザ数分布

とが分かる。また、図３左のフェイクニュースへの信頼度分布を見ると、Pi=0.5～0.7 すなわちフェイクニュースを信じてはいるが、思い込みは強くない人が一定数存在することから、彼らに真実を伝えることで，効果的に説得できる可能性がある。フェイクニュース対策の最も困難な課題である「事実を指摘しても受け入れない」問題に正面から取り組んだ研究は、研究代表者らが知る限り、本研究が初めてである。

本研究の推進中に、「事実を指摘しても受け入れない」問題に対応する別の手法を着想した。すなわち、ニュースをリアルとフェイクに二分し、ユーザを信頼できる人とできない人に二分することは多くの場合非現実的であることが実データの分析から明らかになった。それよりも、立場の異なる２種類のニュースとそれを信じる２つのユーザクラスタとみなす方が自然である。提案法では、相いれない２種類のニュースを発信・転送したユーザアカウント、それらのアカウントが発した投稿(当該フェイクニュース、リアルニュースだけでなく全ての投稿)を収集する。ユーザと単語の関係（例えばあるユーザの投稿中にある単語が出現した回数）をデータ化し、それに共クラスタリングを適用することで、たとえばサッカーに関わる単語集合とそれらの単語を多用するユーザ集合から成るサッカーというトピックを抽出できる。トピックに２つのユーザグループが混在していれば、２つのグループの共通のトピックであり、ユーザの分断を緩和する手掛かりとなる。この手法の実現に向けて、フェイクニュース 11 件、リアルニュース 10 件を発信・転送した Twitter ユーザ 10,438 人が発した全ての投稿 31,645,252 件を収集した。また、ユーザと単語の関係行列を生成するプログラムを開発し、共クラスタリングのライブラリーを入手、試用した。今後は提案法による共通トピック抽出の実装、評価を行いたい。

(3) フェイクニュースの拡散予測

Pryymak モデルに加え、別種の情報共有モデルである Tsang らのモデルを用い、性質の異なる２つの代表的なモデル上で、インフルエンサーに成りすましてフェイクニュースを発信する攻撃をシミュレーションした。なりすましの方法として、インフルエンサーのアカウントを模倣する方法、インフルエンサーのアカウントを乗っ取る方法の 2 種類を評価した。インフルエンサー毎に[0, 1]のランダムな値を生成し、乗っ取りの成功確率とした。平均すると乗っ取りの成功確率は 0.5 とした。評価には、Stanford 大学が公開している Facebook データと Twitter データを用いた。乗っ取りが発覚するまでの時間は先行研究を参考に 10 時間と 20 時間の 2 通りとした。

図 4 に Pryymak モデルを用いた評価結果を示す。図 4 は、模倣によりインフルエンサーに成りすましてフェイクニュースを流すと、80%のユーザはニュースの真偽を判断できなくなることを示している。また、本実験のパラメータの場合は、模倣は乗っ取りよりも強い攻撃である。Tsang らのモデルでも同様の結果を得た。性質の異なる代表的な情報共有モデル上で、同様の結

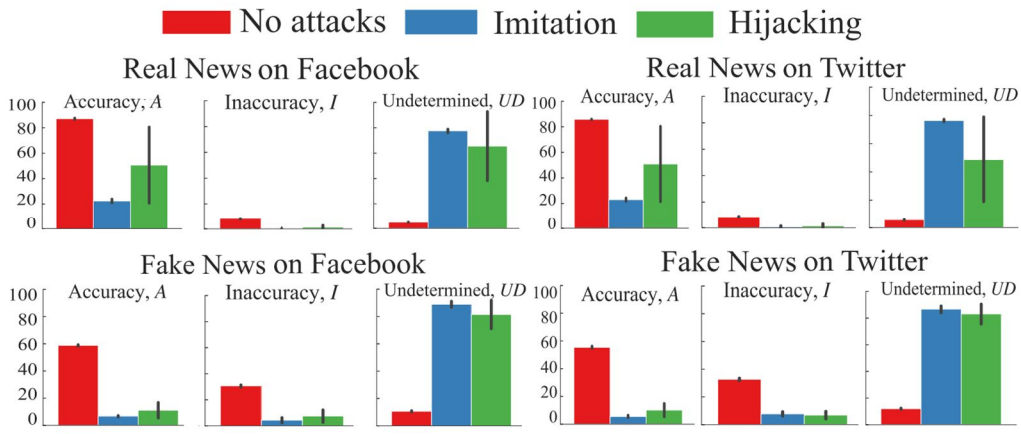

図４　インフルエンサーへのなりすましによるフェイクニュース拡散の効果

果になったことから、これらの結果は一定の信頼性があると考えられる。
Glinton らのモデル、Tsang らのモデルに加えて、さらに異なる性質を持つ Kozma らのモデルを用い、フェイクニュースの発信者が自らの信頼度を実際よりも高く偽装した上でニュースを発信する状況をシミュレーションした。ニュース発信者の信頼度に応じてベイズ更新によりニュースの信頼度を更新する Glinton らのモデルの場合、発信者の信頼度が低くても、受信者が発信者の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、信頼度の低い発信者が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、発信者の信頼度が低くても、受信者が信頼度を正しく把握していれば、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が、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。Tsang らのモデルの場合、発信者の信頼度を高く偽ることで、全体の意見を当該発信者の意見に向けて、早く収束させることができた。以上により、発信者の信頼度が低いことよりも、低い信頼度を高く偽る方が大きな悪影響を与えることを明らかにした。
フェイクニュースの拡散、特に、悪意による計画的な拡散を予測する研究は初めてである。

(4) 画像を用いたフェイクニュースの投稿者特定
研究代表者らは、従来から、ソーシャルメディアの匿名アカウントの所有者を特定する研究を行っていた。その手法は、匿名アカウントの投稿文から単語とその使用頻度を抽出し、それらの特徴量からアカウント所有者の性別、年代、住所、趣味などのプロファイリングを行い、既知の候補者のプロファイルと照合していた。この従来手法を拡張し、投稿画像を用いてアカウントの所有者を特定する手法を検討した。Google 社が公開している Vision API を用いて、投稿画像から単語を抽出し、これらの単語からアカウント所有者の趣味をプロファイリングした。新手法と従来手法を特徴量レベルで統合して、候補者のプロファイルと照合した。51 人の被験者とその Twitter アカウントを用いた照合実験を行い、従来手法のみの場合、新手法による 15 種類の趣味のプロファイリング結果を従来手法と統合する場合を比較した結果、従来手法のみの場合は 51 アカウント中 27 アカウントの所有者を正しく特定できたが、併用法では 31 アカウントを正しく特定できた。これにより、投稿画像を利用して匿名アカウントの所有者すなわちフェイクニュースの投稿者を特定する精度を向上できることが明らかになった。

(5) LLM および生成系 AI を用いて作成されたフェイクニュースへの対策、および対策における LLM および生成系 AI の活用
BERT を用いて投稿文から特徴量を抽出し、ResNet を用いて投稿画像から特徴量を抽出し、それらを接合した特徴ベクトルから機械学習モデル（Support Vector Machine）により、生成系 AI の投稿と人手の投稿を識別する手法を提案した。GPT-4 の文章生成と Stable Diffusion XL の画像生成を組み合わせた生成系 AI の投稿文 4,207 件および、人手による投稿文 12,714 件を用いて評価実験を行った結果、Accuracy が 99.9%であった。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　5件）

１．発表者名
Tomoaki Ohkawa, Hiroshi Yoshiura, Takayasu Yamaguchi

２．発表標題
Explainable Multimodal Fake Posts Detection Using Feature Extraction with Attention Mechanisms

３．学会等名
IEEE International Workshop in Cyber Forensics, Security, and E-discovery (CFSE 2023)（国際学会）

４．発表年
2023年

１．発表者名
Eina Hashimoto, Masatsugu Ichino, Hiroshi Yoshiura

２．発表標題
Breaking Anonymity of Social Media by Profiling from Multimodal Information

３．学会等名
IEEE International Workshop in Cyber Forensics, Security, and E-discovery (CFSE 2022)（国際学会）

４．発表年
2022年

１．発表者名
Kento Yoshikawa, Masatsugu Ichino, Hiroshi Yoshiura

２．発表標題
Modeling Malicious Behaviors and Fake News Dissemination on Social Networks

３．学会等名
IFIP Conference on e-Business, e-Services and e-Society (I3E2021)（国際学会）

４．発表年
2021年

１．発表者名
Risa Kusano, Kento Yoshikawa, Hiroyuki Sato, Masatsugu Ichino, Hiroshi Yoshiura

２．発表標題
Maintaining Soundness of Social Network by Understanding Fake News Dissemination and People's Belief

３．学会等名
International Workshop on Informatics (IWIN2021)（国際学会）

４．発表年
2021年

１．発表者名 Kento Yoshikawa, Takumi Awa, Risa Kusano, Masatsugu Ichino, Hiroshi Yoshiura
２．発表標題 Reliability-Disguised Attacks on Social Network to Accelerate Fake News Dissemination
３．学会等名 IEEE International Workshop in Cyber Forensics, Security, and E-discovery (CFSE 2021)（国際学会）
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	佐藤　寛之 (Sato Hiroyuki) (60550978)	電気通信大学・大学院情報理工学研究科・教授 (12612)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関