ソーシャルメディア上のテキスト情報を考慮したインフルエンサー検出モデル

概要

Twitter や Facebook といった現代のソーシャルネットワーク上には、ユーザー間のつながりを表すネットワーク情報だけでなく、ユーザーの多様な趣味や境遇を表すテキスト情報が存在しており、それらを適切にモデリングした社会ネットワーク分析を行うことが望ましい。本研究では、トピックモデルを用いて両者の情報を考慮したネットワークモデルを構築し、ネットワークの中心的な人物であるインフルエンサーを検出する手法を提案する。先行研究と比較すると、未観測リンクに対しても潜在的な影響力を考慮している点、趣味や境遇が似ていない人同士のリンクであっても一律でない関係性を考慮している点、ユーザーの特徴としてテキスト情報をモデルに組み込んでいる点が新しい。Twitter データを用いた実証分析では、「影響力を持った人からのリンクを集める人はより影響力を持つ」と仮定してインフルエンサーを検出する場合は、提案モデルが既存ネットワークモデルよりも効果的なインフルエンサーを見つけ出すことが出来るという示唆を得ている。

（キーワード：社会ネットワーク分析、ソーシャルメディア分析、インフルエンサー、トピックモデル、ベイズ推定）

1　はじめに

人々は、新製品の採用や情報の拡散を行う際に、友人や家族といった周囲の人々の影響を受けることがある（Reingen et al. 1984）。したがって、商品や情報の普及を成功させるためには、社会ネットワークがどのような構造になっているのかを把握し、その構成員から強い影響力を持った人々を見つけ出すことが重要となっている（Van den Bult and Wuyts 2008）。そして現代は、Twitter や Facebook といったソーシャルメディアが隆盛する時代であり、人々はオフラインだけでなくソーシャルメディア上でも多くの人々とつながり、社会ネットワークを形成している（情報通信白書 2018）。このような現状において、ソーシャルメディア上の社会ネットワークを分析・把握することは、企業のマーケティング活動に必要不可欠なものとなっている。

一方、学術分野では、社会学やマーケティングの分野を中心に多くの研究者が社会ネットワーク分析の研究を行ってきた（Granovetter 1973、Hu and Van den Bult 2014、Aral and Walker 2014）。この社会ネットワーク分析における最大の関心事の一つが、人々の関係性をどのように表現するかである。最も単純で、幅広い領域の問題解決に適用されているのが二値ネットワークの考え方であり、友人関係にあるか否か、メールのやり取りをしたことがあるか否かなど、何らかの関係があればそのリンクは 1 の値を、なければ 0 の値を取ると仮定して社会ネットワークをモデル化している。しかし、Granovetter (1973) を始めとする「人々の結びつきには強弱がある」とする立場に立つと、この二値ネットワークが置く仮定はやや強すぎるものであり、多くの研究者は重み付き連続値ネットワークの考え方を提唱している。

重み付き連続値ネットワークでは、リンクごとに“重み”が異なるため、人々の関係性の強弱を反映したより柔軟なモデリングが可能となっている。そして、その重みは、対象となる二者の関係性を示すリンク情報だけでなく、それぞれがどのような人物であるかを示すノードの特徴にも依存すると考えられている。例えば、実験機器の普及を分析した Hu and Van den Bult (2014) の研究では、過去にその機器を使用した論文をいくつ出しているか、分野はどこか、引用数はいくつかといった情報がノードの特徴として用いられている。一方で、本研究の舞台となるソーシャルメディア上では、多様な趣味や境遇を持ったユーザーが集まっており、どのような関心を持っているのか、あるいはどのようなトピックを投稿しているのかといった、ソーシャルメディア上に投稿されるテキスト情報もノードごとの特徴として使われるべきである。しかし、そのような研究はまだ数が少なく、Aral and Walker (2014) が二者間で共通している Facebook ファンページの数をノードごとの特徴として用いている程度である。ただし、ここで用いられているユーザーの関心を表す情報は定量的なものにとどまっており、より情報量の多い定性的なテキストデータを用いた研究はまだ存在しない。したがって本研究では、ユーザーの関心を表すテキスト情報を考慮した社会ネットワーク分析モデルを提案し、社会ネットワークの把握、及び強い影響力を持ったインフルエンサーの発見を可能にすることを主たる目的とする。

以下、本論文では、2 節で社会ネットワーク分析に関連する先行研究をまとめ、本研究の目的を明確にする。3 節では提案モデルを説明する。次いで 4 節では、Twitter データを利用した実証研究を報告し、5 節で当該ネットワークにおける情報拡散過程のシミュレーション分析を行う。最後に、6 節で結論と今後の課題を述べる。

2　先行研究

2.1　社会ネットワークにおける人々の結びつき

Granovetter (1973) は、ネットワークにおける人々の結びつきには強弱があり、その強さが情報や商品の拡散に大きな影響を与えていると主張した。この主張をきっかけに、社会学やマーケティングの分野を中心として、社会ネットワークにおける結びつきの強さが人々の購買行動に与える影響を明らかにするための研究が行われている。例えば、Moschis and Moore (1979) は、商品を購入するときと価格や性能の評価をするときとでは、影響を受ける人物との関係性が異なることを示している。また、Reingen et al. (1984) は、ルームメイトや勉強のパートナーといった社会的な関係にある人々が、消費者のブランド選択に影響を与えていることを明らかにしている。このように、当初はオフラインでの関係性についての研究が主であったが、近年はソーシャルメディアの流行も相まって、オンラインにおける関係性が盛んに研究されている。例えば Aral and Walker (2014) は、Facebook 上で共通している友人の数などのオンライン上での関係性が商品の採用に影響を与えていると結論付けている。このような結びつきの“強さ”が影響を与えていると主張する研究がある一方で、Godes and Mayzlin (2009) が、友人よりも遠い存在である知人からの口コミが、売り上げに強い影響を与えることを示したように、結びつきの“弱さ”が重要な要素であると主張する研究もある。

2.2　社会ネットワーク分析研究の変化と課題

2.2.1　社会ネットワーク分析研究の変化

このように社会的な関係性には強弱があり、それはリンク毎に異なるものだとする研究結果が増えてきたことを受けて、社会ネットワーク分析におけるネットワークの捉え方が現在進行形で変化している。まず、最も単純で強い仮定を必要とするものが二値ネットワークである。二値ネットワークにおいては、ある二者間に着目したときに、彼らが友人であるか否か（Park et al. 2018）、彼らの間でメールのやり取りがあったか否か（Hinz et al. 2011）など、何らかの関係性があればそのリンクは 1 を、なければ 0 の値を取ることとなる。しかし、この考え方は上述した結びつきの強さの議論から考えるとやや不適切であり、強すぎる仮定として批判されている（Chen et al. 2017）。そこで提案された次なる考え方が二値ネットワークを複数組み合わせるという考え方である（Aral and Walker 2014、Hu and Vanden Bult 2014、Iyengar et al. 2011）。Iyengar et al. (2011) では、内科医に対してアンケートを行い、症状に関して共に議論したい医者と患者を紹介したい医者をそれぞれ複数人挙げてもらうことで 2 種類の二値ネットワークを構築し、それらを足し合わせた“Total”ネットワークによる分析を行っている。しかし、これら二値ネットワークも複数二値ネットワーク（あるいは多値ネットワーク）も、ネットワークを形成するための重みは離散値であり、ここにも強い仮定があるとして重みを連続値に拡張した、重み付き連続値ネットワークが提案されている（Trusov et al. 2010、Ansari et al. 2011、Chen et al. 2017）。重み付き連続値ネットワークにおけるリンクの重みは、リンクごとに異なる連続値で表されるため、人々の関係性をより柔軟にモデリングすることが出来る。Chen et al. (2017) は、インドのある村を対象に、世帯間の経済的社会的なつながりなどを調べて複数の二値ネットワークを構築し、それらの加重和を取ることで重み付き連続値ネットワークを構築している。

2.2.2　社会ネットワーク分析研究の課題（１）：未観測リンクの扱い

本研究もまた、上述した重み付き連続値ネットワークの構築を目指す研究ではあるが、先行研究で解決されていない課題として、本研究で問題提起するものの一つが未観測リンクの扱いである。リンクが観測された関係性については、上述したように、二値や連続値などいくつかの扱い方がある。しかし、リンクが観測されていない関係性、例えば友人ではない間柄や、ソーシャルメディアでフォロー・フォロワーの関係にはない二人のリンクなどについては、多くの研究で一律に 0 を与えている。しかし、実際にはそのような関係にも潜在的な、あるいは将来的な関係性や影響力が存在しているはずである。友人関係ではない二人であったとしても、まだ（深い）知り合いになっていないだけで、共通の趣味や境遇をきっかけに将来的に影響を与え合う関係になる可能性がある。また、実務的な視点から見ると、アンケートなどを基に対象とする人々のネットワークを構築するような手法の場合、関係のある人々を全て数え上げることは難しく、実際に影響力を持っているはずのリンクをデータとして観測できていない可能性がある（Chen et al. 2017 でも今後の課題として未観測リンクの扱いに言及している）。この問題に対する回答の一つが、確率的な潜在変数を仮定したモデルである。Airoldi et al. (2008) は、リンクに潜在的なトピックを仮定した混合メンバシップ確率的ブロックモデル（Mixed Membership Stochastic Blockmodels、以下 MMSB）により、0 リンク問題に取り組んでいる。本研究はこの MMSB の考えに基づくモデリングを行っており（詳しくは 3 節にて後述）、したがって、本研究の第一の目的は、確率的潜在変数モデルにより、未観測リンクに対しても確率的な重みを与える重み付き連続値ネットワークを構築することである。

2.2.3　社会ネットワーク分析研究の課題（２）：似ていない他者とのリンクの可能性

先行研究で解決されていない課題の二つ目は、似ていない他者とのリンクの可能性である。社会学におけるネットワーク理論として同類性（homophily）という考え方があり、それは、類似した特徴を持つ者同士は、そうでない関係性と比べてリンクを結びやすく、相互に影響しあう傾向にあるという理論である。多くの研究がこの理論に基づき、類似した二者間のリンクには強い重みを、そうでないリンクには弱い重みを付している。しかし、類似度に従って一律に重みを付すという仮定は、状況によってはやや不適切な場合もあり、より柔軟なモデリングが必要である。ソーシャルメディアが発達し、世界中の様々な人と社会的なつながりを持てるようになった現代において、共通の趣味や境遇のない、つまり自分と似ていない他者とつながり、その相手から様々な影響を受けるという状況は想像に難くない。また、Godes and Mayzlin (2009) は、弱い関係（知人）の方が、強い関係（友人）よりも強い影響力を持っているという“弱い紐帯の強さ”理論を唱えている。一般的に考えれば、強い関係よりも弱い関係にある人との方が類似していないはずである。よって、強い影響力を持っている可能性があるにもかかわらず、類似していない者同士のリンクに一律に弱い重みを与えるというモデリングは不適切であると分かる。この問題に対処するために、例えば Braun and Bonfer (2011) は、観測されたリンクには類似した関係と類似していない関係が混在していると仮定し、観測ネットワークを潜在空間に射影することで、真に類似した関係性を見つけ出している。本研究で提案するモデルでは、ユーザーがそれぞれ持つ潜在的なトピック分布と、そのトピック同士が結びつくリンク確率を推定することで、ユーザー同士のトピック分布が似ていても似ていなくても、リンク確率に従った一律でない重みをリンクに与えることが出来る。したがって、本研究の二つ目の目的は、似ていない人同士であっても大きな影響力を持つ可能性があることを考慮したネットワークを構築することである。

2.2.4　社会ネットワーク分析研究の課題（3）：ソーシャルメディア上のテキスト情報の利用

本研究で問題提起する最後の課題は、ソーシャルメディア上のテキスト情報の利用である。先行研究では、ネットワークの重みを決めるために、人々のつながりを表すリンク情報（共通の友人の数：Aral and Walker 2014、論文の共著者であるか否か：Hu and Van den Bult 2014、経済的なつながり：Chen et al. 2017 など）だけでなく、ノードごとの特徴（過去のログイン情報：Trusov et al. 2010、特定機器を使用した論文の数：Hu and Van den Bult 2014 など）が用いられている。本研究ではこのノードごとの特徴としてソーシャルメディア上に投稿されるテキスト情報、あるいはその背後にあるユーザーが持つ関心事を利用するべきであると考える。ソーシャルメディア上のネットワークは、他の社会的なネットワーク（医者のコミュニティ：Iyengar et al. 2011、ゲームユーザーのコミュニティ：Park et al. 2018 など）と異なり、多様な趣味や境遇を持ったユーザーが集まっている。そのような人々の間に結ばれるリンクは、相手がどのような人とつながっているかなどのリンク情報は勿論のこと、相手がどのような関心を持っているのか、あるいはどのようなトピックを投稿しているのかといったテキスト情報が与える影響も小さくないと考えるべきである。しかし、このようなソーシャルメディア上のテキスト情報、ないしはユーザーの関心をモデルに組み込んだネットワーク分析の研究は数が少なく、Aral and Walker (2014) が、Facebook のファンページをネットワークの重みを決めるための変数として加えることで、ユーザーの関心を考慮している程度である。本研究で提案するモデルは、次節で詳しく述べるが、テキスト情報とネットワーク情報が相互に影響を与えながらトピックを推定する構造を持っており、両者を考慮したネットワークの重みづけが可能となっている。したがって、本研究の目的の最後は、ネットワーク情報に加えてユーザーの特徴を表すテキスト情報を考慮したネットワークを構築することである。

以上のように、本項では先行研究で残された課題三点を提起した。それらは、「未観測リンクの扱い」「似ていない他者とのリンクの可能性」「ソーシャルメディア上のテキスト情報の利用」である。これらの課題を解決するべく、次節では、潜在的なトピックを仮定したモデルと、そこから得られるパラメータを用いた重み付きネットワークモデルを提案する。

3　モデル

3.1　3 つのトピックモデル

3.1.1　潜在的ディリクレ配分法

本項では、提案モデルを議論するうえで基礎となる 3 種類のトピックモデルについて説明する [1]。まず一つ目は、潜在的ディリクレ配分法（Latent Dirichlet Allocation、Blei et al. 2003、以下 LDA）である。LDA は、文書データのためのモデリングであり、一つの文書が複数のトピックを持つことを仮定する。例えば、「オリンピックの経済効果」に関する新聞記事であれば、「スポーツ」と「経済」という 2 つのトピックを持っていると考えられる。LDA では、これを文書ごとに異なるトピック分布 $\theta_d = (\theta_{d1}, \ldots, \theta_{dK})$ で表現する（新聞記事が持つトピック分布は、スポーツと経済のトピックで高い値をとる）。ここで、$\theta_{dk} = p(k|\theta_d)$ は文書 d 内の単語にトピック k が割り当てられる確率を表し、$\theta_{dk} \geq 0, \Sigma_k \theta_{dk} = 1$ を満たす。また、LDA はトピック毎に異なる単語分布 $\phi_k = (\phi_{k1}, \ldots, \phi_{kV})$ を持っており、$\phi_{kv} = p(v|\phi_k)$ は、トピック k において語彙 v が出現する確率を表し、$\phi_{kv} \geq 0, \Sigma_v \phi_{kv} = 1$ を満たす。そして、トピック分布 θ_d に従って文書 d 内の n 番目の単語にトピック z_{dn} が割り当てられ、割り当てられたトピックに対応する単語分布 $\phi_{z_{dn}}$ に従って単語 w_{dn} が生成されるというプロセスが、

単語の生成過程としてモデリングされている。

$$z_{dn} \sim Categorical\left(\theta_d\right), \quad w_{dn} \sim \left(\phi_{z_{dn}}\right) \tag{1}$$

3.1.2 混合メンバシップ確率的ブロックモデル

次に説明するモデルは、混合メンバシップ確率的ブロックモデル（Mixed Membership Stochastic Blockmodel、Airoldi et al. 2008、以下 MMSB）である。MMSB は、ネットワーク上のリンクに潜在的なトピックを仮定するモデルであり、LDA が「一つの文書が複数のトピックを持つ」状況を想定していたのに対し、MMSB は「一つのノードが複数のトピックを持つ」状況を仮定している。そのため、LDA における文書と単語の関係性は、MMSB におけるノードとリンクに対応している。また、モデルの構造も LDA と対応させて考えることができ、MMSB は、ノードごとに異なるトピック分布 θ_d と、トピック同士がネットワーク上でリンクする確率 $\phi = \{\phi_{kk'}\}$ をパラメータとして持つ。ただし、一本のリンクに対して、送り手（Sender）側と受け手（Receiver）側の二つのトピックが割り当てられる点は、LDA における単語トピックの割り当てとは異なる。つまり、ノード d からノード d' へのリンクであれば、リンクの送り手側のトピック $S_{dd'}$ は θ_d に、受け手側のトピック $R_{dd'}$ は $\theta_{d'}$ にしたがって割り当てられ、その割り当てられたトピックに対応するリンク確率 $\phi_{S_{dd'}R_{dd'}}$ に従ってリンク $y_{dd'}$（リンクがあれば 1 を、なければ 0 を取る）が生成される。

$$S_{dd'} \sim Categorical\left(\theta_d\right), \quad R_{dd'} \sim Categorical\left(\theta_{d'}\right), \quad y_{dd'} \sim Binomial\left(\phi_{S_{dd'}R_{dd'}}\right) \tag{2}$$

LDA が単語の潜在的な意味を捉えるモデルであったのに対し、MMSB はリンクの潜在的なクラスタを捉えたモデルと言えるため、社会学などの分野では、しばしばコミュニティの検出を目的として MMSB が用いられている（Sweet and Zheng 2018）。

3.1.3 対応トピックモデル

最後に説明するトピックモデルは対応トピックモデル（Correspondence Topic Model、Blei and Jordan 2003、以下 CTM）である。上述した 2 つのモデルが文書内のテキストやネットワーク上のリンクといった単一の情報を扱っていたのに対して、CTM では複数の情報を対応付けてモデリングを行う。例えば、インターネット上のブログ記事には、本文のテキスト情報に加えて、検索を容易にするためのタグが付いている。このタグは、本文の内容と連動して付けられるため、本文の潜在トピックを推定する際に使用する補助的な情報として捉えることができる。このとき、CTM では、その文書内の単語に割り当てられたトピックの割合 $\frac{N_{dk}}{N_d}$ に応じて補助情報トピック y_{dm} が割り当てられ、そのトピックに対応する補助情報分布 $\psi_{y_{dm}}$ に従って補助情報 x_{dm} が生成される。

$$y_{dm} \sim Categorical\left(\frac{N_{d1}}{N_d}, \ldots, \frac{N_{dK}}{N_d}\right), \quad x_{dm} \sim Categorical\left(\psi_{y_{dm}}\right) \tag{3}$$

ただし、N_{dk} は文書 d 内でトピック k が割り当てられた単語の数、N_d は文書 d 内の単語の数を表す。比較のため、ここで説明した 3 種類のトピックモデルについて、そのグラフィカルモデルを図 1 に示す。

このように、3 種類のトピックモデルを説明したが、次項で説明する提案モデルはこれらのモデルが基礎となっている。前節でも述べたように、本研究では、ユーザー同士のつながりを表すネットワーク情報と各ユーザーが投稿するテキスト情報とを考慮したネットワーク構築を目的としているため、ネットワーク情報については MMSB の、テキスト情報については LDA の考え方に基づいてモデリングを行い、両者のトピック対応関係を CTM に従って反映させたモデルを提案する。

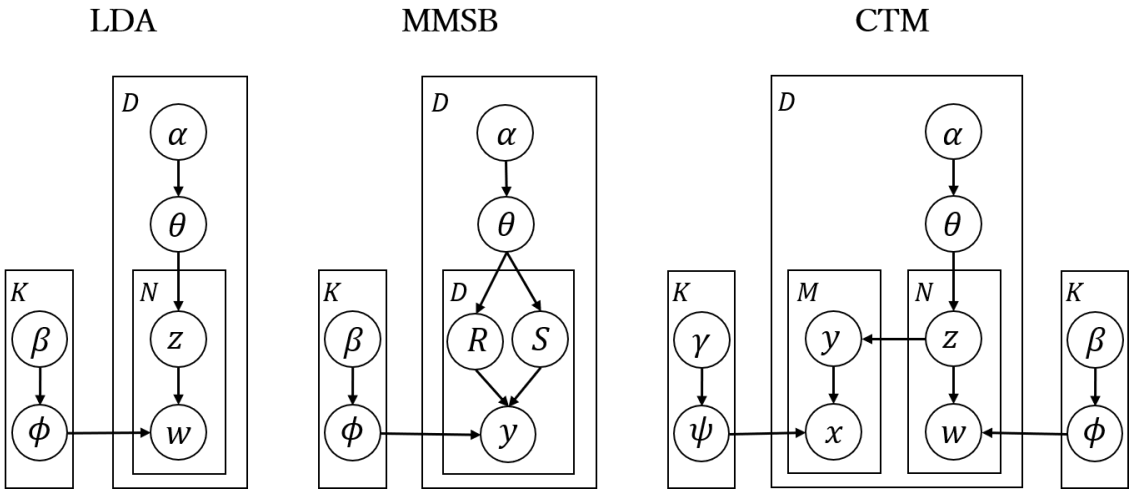

図 1　3 つのトピックモデルのグラフィカル表現

3.2　提案モデル

3.2.1　提案トピックモデル

まず、本研究で提案するトピックモデルにおいて、リンク及び単語にトピックが割り当てられる過程を説明する。ネットワーク上の各ユーザーは、それぞれに異なるトピック分布 θ_d を持ち、それに従って各リンクにトピック $S_{dd'}, R_{dd'}$ が割り当てられる。そして、ユーザー d に関係するリンクに割り当てられたトピックの割合に応じて、ユーザー d が投稿するトピック z_{dn} が決まる。

$$\begin{aligned} S_{dd'} \sim Categorical\left(\theta_d\right), \quad R_{dd'} \sim Categorical\left(\theta_{d'}\right), \\ z_{dn} \sim Categorical\left(\frac{N_{d1}}{2(D-1)}, \ldots, \frac{N_{dK}}{2(D-1)}\right) \end{aligned} \tag{4}$$

ただし、N_{dk} は、ユーザー d が関係するリンクのうちトピック k が割り当てられた数である。

次に、リンク及び単語の生成過程を説明する。提案モデルでは、LDA 同様、トピック毎に異なる単語分布 ϕ_k を持つが、リンク確率 $\psi^{(d)} = \{\psi^{(d)}_{kk'}\}$ には、MMSB のように全ユーザー共通ではなく、ユーザーごとに異なるという仮定を置いている。このように、リンク確率に異質性を導入することで、同じトピックを持つリンクであっても、そのリンクが関係するユーザーの社会的影響力などによって結ばれる確率が異なるという、より現実に即したモデルとなっている。そして、この 2 つのパラメータと割り当てられたトピックに応じてリンク及び単語が生成される。

$$y_{dd'} \sim Binomial\left(\psi^{(d')}_{S_{dd'}R_{dd'}}\right), \quad w_{dn} \sim Categorical\left(\phi_{z_{dn}}\right) \tag{5}$$

図 2 は、提案トピックモデルのグラフィカルモデルであり、モデル全体の生成過程と各パラメータの事前分布は Appendix に示す。

3.2.2　重み付き連続値ネットワークの構築

前節では、重み付き連続値ネットワークを構築するうえで、解決すべき課題を挙げたが、本研究で提案するモデルから得られるパラメータによって、それらが達成されうることを本項で示す。

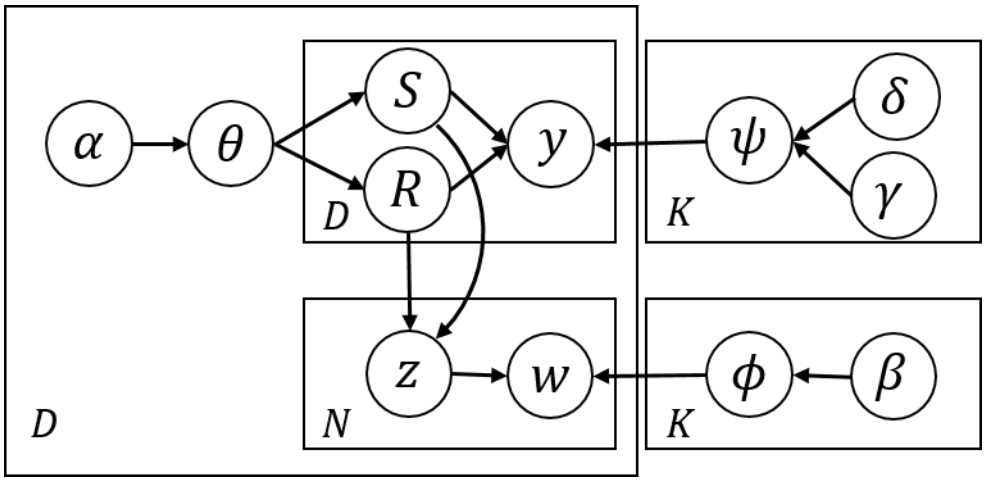

図 2　提案トピックモデルのグラフィカル表現

提案モデル（及び MMSB）は、0 か 1 で表されるネットワークに対して潜在的なトピックを仮定することで、そのリンクが生成される背後にある連続的な確率を推定するモデルであった。本研究では、この推定された確率を二者間の関係性の強さと捉えることで、リンクの重みを表現する。ユーザー d から d' へのリンクを考えたときに、このリンクが結ばれる確率が高い、つまりユーザー d が d' をフォローしたいと強く思っているときほど、二者間の関係性が強く、ユーザー d' が発信した情報がユーザー d に大きな影響を与えると考えるのである。

提案モデルにおいて、d から d' へのリンクの生成に関するパラメータであるトピック分布 $\theta_d, \theta_{d'}$ とトピック間のリンク確率 $\psi^{(d')}$ が得られたとき、このリンクが結ばれる確率は、

$$p\left(y_{dd'}=1|\theta_d,\theta_{d'},\psi^{(d')}\right)=\theta_d\psi^{(d')}\theta_{d'}^T=\sum_{k=1}^{K}\sum_{k'=1}^{K}\left(\theta_d^T\theta_{d'}\odot\psi^{(d')}\right)_{kk'} \tag{6}$$

で表される。ただし、$\odot$ はアダマール積を表し、$()_{kk'}$ は、カッコ内の行列における k, k' 成分を意味する。このとき、$\theta_d^T\theta_{d'}$ はユーザー d, d' 間のトピックごとの類似度を表しており、$\theta_d^T\theta_{d'}$ で表される行列の各成分を足し合わせれば全てのトピックを考慮した類似度となる。つまり本研究では、二者間の関係性をトピックごとの類似度とトピック間のリンク確率の積で表現している。さらに、ユーザー d' が発信する情報がユーザー d に与える影響のうち、トピック k に関する影響（$x_{dd'}^{(k)}$ とする）を知るには、式（6）右辺のカッコ内で表される確率行列を k について周辺化すればよい。

$$x_{dd'}^{(k)}=\sum_{k'=1}^{K}\theta_{dk}\theta_{d'k'}\psi_{kk'}^{(d)} \tag{7}$$

そうすることで、一つのネットワークに対してトピックごとの影響力を把握することが出来るため、企業は、拡散したい情報や商品のトピックに合わせてインフルエンサーを見つけ出すことが出来る。

このようにして、リンクごと・トピックごとに重みが連続的に異なるネットワークを構築することができたため、たとえ未観測のリンクであっても重みを与えることができる。また、似ていないユーザー同士、つまり両者のトピック分布が大きく異なるユーザー同士であっても、リンク確率に従って一律でない重みを与えることができる。そして、前項でも説明したように、ネットワーク情報とテキスト情報を対応付けてトピックが決まるため、リンクごとの重みも、ユーザーが投稿するテキスト情報を反映し

たものとなっている（トピックの推定に両者の情報が使われていることは、Appendix に示すギブスサンプリングのサンプリング式からも見て取れる）。

ここで説明したような、性質の異なる複数のネットワークを構築するという考え方は、Iyengar et al. (2011) などの二値ネットワークの研究でも、Hu and Van den Bult (2014) や Chen et al. (2017) などの重み付きネットワークの研究でも用いられている。しかし、これらの研究では、複数のネットワークそれぞれが情報拡散に与える影響をデータから推定し、それらの加重和を取ることで一つのネットワークを構築しているのに対し、本研究では、トピックの数だけのネットワークを構築するにとどまっている。本研究を応用してネットワークが持つ情報拡散への影響を把握するには、拡散過程を追跡したデータが必要であり、これは今後の課題である。

3.2.3 インフルエンサーの検出

ここでは、前項でモデリングした重み付きネットワークからインフルエンサーを検出する手法を説明する。社会学を中心としたグラフ理論の研究において、ネットワーク上で重要なノードを発見する手法として中心性という概念がある（Freeman 1979、Buckley and Harary 1990）。本研究でもその考え方に即して、次数中心性（degree centrality）と固有ベクトル中心性（eigen vector centrality）の二つの指標を用いたインフルエンサーの検出を行う。

次数中心性とは、あるノードが関係するリンクの重みの和で表される指標であり、トピック k に関する影響力で重みづけされたネットワーク $x^{(k)} = \{x^{(k)}_{dd'}\}$ が与えられたとき、ユーザー d の次数中心性は以下の式で定義される。

$$deg^{(k)}_{d} = \sum_{d'=1}^{D} x^{(k)}_{d'd} \tag{8}$$

次数中心性の値が高いユーザーほどインフルエンサーとしての素質を持っているとみなすため、次数中心性によって検出されるインフルエンサーは、影響力を持ったリンクを多く集めるユーザーであると言える。

もう一つの指標である固有ベクトル中心性では、自分に向かうリンクだけでなく、自分が関係する人々が持つ中心性も考慮される。ネットワーク $x^{(k)}$ に対する固有ベクトルを $v^{(k)} = \{v^{(k)}_{d}\}$ とすると、ユーザー d の固有ベクトル中心性は、固有ベクトル $v^{(k)}$ の第 d 成分で定義される。

$$ev^{(k)}_{d} = v^{(k)}_{d}, \quad x^{(k)}v^{(k)} = \lambda^{x^{(k)}}_{max}v^{(k)} \tag{9}$$

ただし、$\lambda^{x^{(k)}}_{max}$ は、ネットワーク $x^{(k)}$ に対する最大固有値である。こちらも同様に、固有ベクトル中心性の値が高いほどインフルエンサーとしての素質を持っているとみなされる。よって、固有ベクトル中心性によって検出されるインフルエンサーは、「影響力の強いユーザーからのリンクを集めるユーザーはより強い影響力を持つ」という仮定の上に立っていると言える [2]。

Borgatti (2005) は、この二つの中心性を、ネットワーク上の影響力を捉えることに適した指標であると評価しており、Chen et al. (2017) は、両中心性のどちらを用いたモデルが情報拡散のデータにより当てはまるかを検証している。本研究でもこれらの研究を参照し、二つの中心性によって検出される性質の異なるインフルエンサーがどのように情報を拡散するのか、そして二値ネットワークから検出されたインフルエンサーと重み付きネットワークから検出されたインフルエンサーが異なる情報拡散能力を持つのかについて、5 節でシミュレーション分析を行う。

4　実証分析

4.1　データ

本研究では、筆者らが収集した Twitter データを用いて提案モデルの実証分析を行う。対象とするネットワークは、任天堂株式会社のアメリカにおける子会社である、Nintendo of America Inc. が所有する Twitter 公式アカウント（アカウント名：@NintendoAmerica）を中心とするエゴネットワークである。この公式アカウントをフォローするユーザーの中から無作為にユーザーを抽出し（1st-step）、その抽出されたユーザーをフォローするユーザーの中からさらに抽出を行い（2nd-step）、計 5,028 人を対象としている。ただし、一企業が SNS 上で抱えるフォロワーの中からインフルエンサーを見つけ出すという状況を想定すれば、ユーザーの無作為抽出を行わず、全てのユーザーを対象にして分析するべきだが、Nintendo of America 公式アカウントのフォロワーは 900 万人を超えており [3]、提案モデルではとても扱いきれないデータ量になってしまう。推定法の改善も含めた、本研究の大規模データへの応用は今後の課題としたい。

これらの対象ユーザーについて、全ユーザー間のフォロー関係、及び自身のタイムライン上に投稿したテキストデータを収集した。さらに、テキストデータは、分析で用いるにあたっていくつかの前処理を施しており、それらの詳細なステップは Appendix に記載している。収集したデータの概要は表 1 に示している。また、今回収集したデータは、ユーザー間のリンク関係については 2018 年 5 月 1 日時点のものを、テキストデータについては 2017 年 9 月 1 日から 2018 年 2 月 28 日までに投稿されたものを対象としており [4]、両者の測定期間に開きがあるため、テキストデータの対象期間からネットワークの対象時点まで、ユーザー間のリンク関係が大きく変動していないことが前提となっていることに注意されたい。

表 1　Twitter データの概要

	ユーザー数	フォロワー数				語彙数				
		最大値	最小値	平均	標準偏差	最大値	最小値	平均	標準偏差	総数
全体	5,028	744	0	15.08	35.94	48	1	10.77	7.02	19,805
1st-step	887	744	0	25.45	54.73	46	2	13.04	6.78	-
2nd-step	4,141	508	0	12.86	29.98	48	1	10.29	6.98	-

4.2　推定結果

4.2.1　Perplexity によるモデル比較

パラメータの推定には、崩壊型ギブスサンプリングを用いており、1,000 回の繰り返しを行い、最後の繰り返しで得られたトピック割り当てを各リンク及び各単語の推定トピックとしている [5]。なお、ユーザーごとランダムに、90% のリンクと単語を訓練データとして用いており、残りの 10 %をテストデータとして Perplexity の計算と次項で示す予測精度を測るためのサンプルに使用している。また、トピック数は、5・10・15・20・25 の候補の中から Perplexity が最良（小さいほど良い）となるものを選んでおり、表 2 のような結果が得られたため、これ以降はトピック数を 15 としたモデルによる推定結果の解釈を行う。崩壊型ギブスサンプリングの詳細なアルゴリズムと Perplexity の定義は Appendix に示す。

表 2　Perplexity によるモデル比較

トピック数	5	10	15	20	25
Perplexity	938.39	799.15	**765.95**	791.31	802.45

4.2.2　トピック分布の解釈

まず、推定されたトピック分布のうち、1 番から 20 番までのユーザーのトピック分布を例として図 3 に示す。トピック分布は、各ユーザーの潜在的な所属コミュニティを表すが、いくつかのトピック分布（No.11, 12, 14 など）は、複数のトピックに対して高い値を取っており、ソフトクラスタリングが行われていることが分かる。また、図 4 は、トピック毎に全ユーザーのトピック分布を足し合わせ、その割合を棒グラフにしたものである。トピック 6 が最も多く、次いでトピック 4 と 9 が多いという結果であった。これは、多くのユーザーがこれらのトピックの特徴を持っていることを表しているため、トピックの解釈を行う際の優先順位の目安となる。

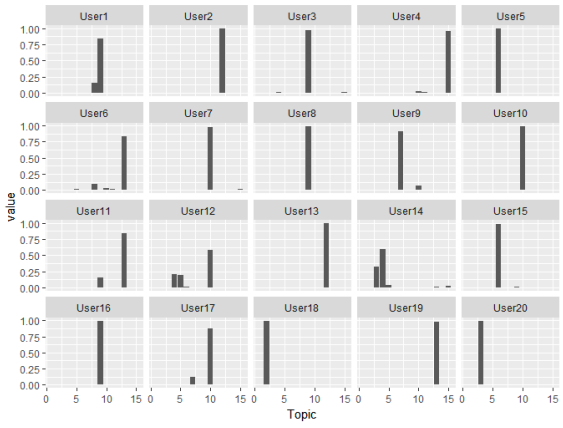

図 3　ユーザーごとのトピック分布

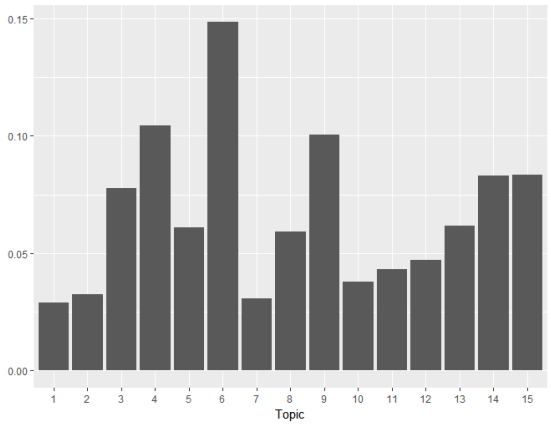

図 4　トピック分布の割合

4.2.3　リンク確率の解釈

次に、推定されたリンク確率のうち、1 番と 1319 番のユーザーのリンク確率を例として表 3 に示す（紙面上の都合によりトピック 1・8・9 のみ抜粋）。各列の左端の値がリンク確率の推定値であり、行に示すトピックから列に示すトピックへのリンクが結ばれる確率を表す。隣の 2 種類の数字は、そのトピック同士のリンク関係のうち、実際に結ばれている数（左）と結ばれていない数（右）のカウントである。これを見ると、ユーザー 1 は、受け手側としてトピック 8 と 9 が割り当てられたリンクを多く持っている（これは図 3 のトピック分布からも分かる）が、そのほとんどが未観測リンクであり、そのためリンク確率も全体的に低いということが分かる [6]。一方で、ユーザー 1319 は、受け手側としてのリンクのほとんどにトピック 1 が割り当たっており、同じトピック 1 を持つユーザーからのリンクだけでなく、他のトピック（トピック 2・7・11 など）を持つユーザーからのリンクに対しても高い確率で結ばれている。このように、ユーザーごとにリンク確率を異なるとすることで、影響力の強さを表現することが出来る。

表 3　リンク確率

	ユーザー 1								ユーザー 1319							
	Topic 1		～	Topic 8		Topic 9		～	Topic 1		～	Topic 8		Topic 9		～
	推定値	カウント		推定値	カウント	推定値	カウント		推定値	カウント		推定値	カウント	推定値	カウント	
Topic 1	0.67	0 0		0.04	0 21	0.01	0 118		0.26	34 102		0.25	0 1	0.33	0 0	
Topic 2	0.33	0 0		0.06	0 13	0.01	0 143		0.16	22 121		0.40	1 1	0.33	0 0	
Topic 3	0.33	0 0		0.02	0 50	0.01	1 303		0.14	47 295		0.40	1 1	0.33	0 0	
Topic 4	0.33	0 0		0.01	0 78	0.00	0 398		0.06	28 438		0.60	2 0	0.33	0 0	
Topic 5	0.33	0 0		0.02	0 43	0.01	1 226		0.06	17 262		0.33	0 0	0.33	0 0	
Topic 6	0.33	0 0		0.03	2 83	0.00	1 578		0.06	40 600		0.50	2 1	0.33	0 0	
Topic 7	0.33	0 0		0.04	0 21	0.01	0 114		0.18	24 112		0.25	0 1	0.33	0 0	
Topic 8	0.33	0 0		0.04	0 45	0.00	0 223		0.12	30 226		0.75	1 0	0.33	0 0	
Topic 9	0.33	0 1		0.01	0 76	0.01	0 391		0.08	35 405		0.20	0 2	0.67	0 0	
Topic 10	0.33	0 0		0.04	0 25	0.01	0 135		0.08	13 161		0.33	0 0	0.33	0 0	
Topic 11	0.33	0 0		0.03	0 30	0.01	0 168		0.22	42 148		0.33	0 0	0.33	0 0	
Topic 12	0.33	0 0		0.03	0 37	0.01	0 187		0.10	19 185		0.40	1 1	0.33	0 0	
Topic 13	0.33	0 0		0.02	0 48	0.00	0 229		0.07	19 258		0.33	0 0	0.33	0 0	
Topic 14	0.33	0 0		0.03	1 56	0.00	0 309		0.08	29 339		0.33	0 0	0.33	0 0	
Topic 15	0.33	0 0		0.02	0 50	0.00	0 319		0.05	18 345		0.33	0 0	0.50	1 0	

4.2.4　単語分布の解釈

提案モデルにおける最後のパラメータである単語分布について解釈する。推定された単語分布の各トピックについて、値の高い上位 10 個の単語を表 4 に示している。例えば、Topic 3 であれば、Twitter 上で個人的にゲーム開発を行っているユーザーが使うハッシュタグである「#gamedevelopment」や、ゲーム実況などに関する投稿で使われる「nowplaying」などの単語が上位に来ており [7]、コアなファンであることを表すトピックであると推察される。また、Topic 7 であれば、「amazon」や「review」といった単語が上位に来ているため、ゲームソフトのレビューに関する投稿よく行うユーザーであることを表すトピックであると推察される。このように、トピックの解釈を行うことで、3.2.2 項で構築したトピックごとのネットワークを解釈することが出来、企業はこのトピックと自らが拡散したい商品や情報とを考慮して、インフルエンサーの検出に取り組むことが出来る。ただし、全てのトピックについて説得力のある解釈が出来ているわけではなく、訓練に用いる単語の洗練や、ラベル付きトピックモデル（Ramage et al. 2009）などによるトピック解釈性の向上などを行うといった課題が残されている。

4.3　予測精度の検証

前項で推定されたモデルを用いて訓練データ・テストデータのリンクを予測することで、モデルのデータに対する当てはまりの良さと未知データに対する予測精度を検証する。3.3 項でも説明したように、ある二者間のリンクが結ばれる確率は式（6）で表される。予測を行うためには、全リンクに対してこの確率を計算し、ある閾値を超えればリンク有、超えていなければリンク無と判定する。このとき、予測精度は設定した閾値に依存することとなるが、閾値を徐々に変化させていった時の感度（True Positive Rate）と偽陽性率（False Positive Rate）のトレードオフの関係性をグラフにしたものは、ROC（Receiver Operating Characteristic）曲線と呼ばれ、図 5・6 に示されている。ROC 曲線の下側の黒く塗られている部分が AUC（Area Under the Curve）であり、これは完全にランダムな分類の時に 0.5 となり、1 に近づくほど良いモデルであるとされる。訓練データ・テストデータにおける AUC の値はそれぞれ 0.882・0.840 であった。

表 4　単語分布における頻出単語 Top 10

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
automat	stat	gamedev	chanc	pop	anim	amazon	travel
gain	indiedev	nowplay	winner	pokemon	irl	review	twitch
gemini	song	gt	exclus	enter	mixtap	digit	stream
trapadr	album	vs	dragon	shini	check	wine	♀
level	daili	morn	ball	walkthrough	meirl	romanc	supportsmallstream
brand	track	wrightwaysradio	£	sampl	charact	bluray	instrument
citi	cupcakenew	sourc	splatoon	grind	cosplay	cupcak	hour
check	bitcoin	espn	sweepstak	year	state	paul	anim
vodanil	indi	horror	star	eagl	scorpio	poppi	capricorn
fun	pixelart	gameplay	prize	wolfpack	smash	esport	streamer

Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
trailer	nintendo	nonfollow	footballnew	xbox	quot	smash
splatoon	mario	blog	presid	ps	startup	pokemon
mario	nintendoswitch	part	indiegam	footbal	sagittarius	soundcloud
youtub	switch	entrepreneur	democrat	playstat	marijuana	post
aquarius	odyssey	busi	tax	heart	forc	ecommerc
movi	cancer	libra	senat	kingdom	tip	pm
war	growthhack	webcam	cork	toonami	fl	tournament
virgo	card	socialmedia	mario	dragon	marijuanalaw	health
leagu	zelda	goal	nintendoswitch	pc	reader	mele
podcast	hero	smfaninja	nintendo	netneutr	minecraft	vs

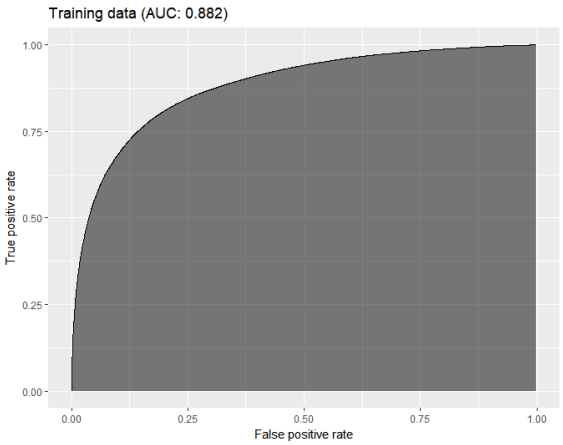

図 5　ROC 曲線（訓練データ）

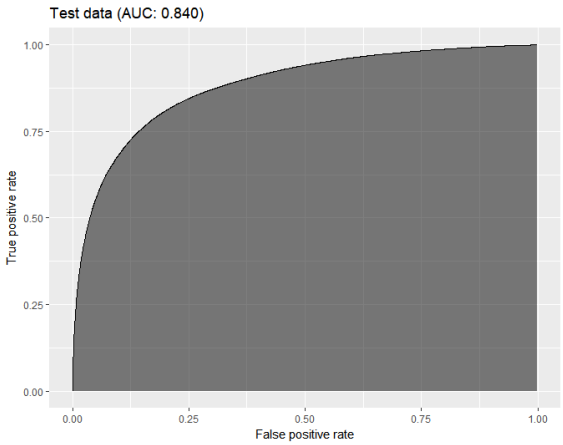

図 6　ROC 曲線（テストデータ）

本研究で用いたデータに限らず、多くのソーシャルネットワークは、少数のリンクのみが結ばれ、大半のリンクが結ばれていない疎なネットワークである。このようなインバランスな分類問題における予測では、極端に言うと、全てを多数派のクラスに予測しただけでも全体の正答率（Accuracy）はそれなりの精度を誇ってしまう。そこで、クラス・インバランスなデータであっても頑健な評価方法が必要となるが、そのような指標として MCC（Matthews Correlation Coefficient、Matthews 1975）があり、次式で定義される。

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

ただし、TP、TN、FP、FN は、それぞれ True Positive、True Negative、False Positive、False Negative

の頭文字である。MCC は、-1 から 1 までの値を取り、完全に正しく予測したときに 1、完全に誤って予測したときに-1、ランダムな予測では 0 となる。表 5 及び 6 は、MCC が最大となるように閾値を設定したときの混同行列である。表 5 では、閾値は 0.0467 であり、MCC 及び正答率を計算すると、それぞれ 0.145、0.987 である。同様に表 6 では、閾値が 0.0553、MCC と正答率が、0.131、0.990 であった。上述した AUC の基準では、訓練データ・テストデータともに 0.8 を超えており、高い性能を持つ予測器であると評価できるが、MCC は 0.1 代と決して高いとは言えず、改善が必要である。これは今後の課題としたい。

表 5　混同行列（訓練データ）

		予測	
		リンク有	リンク無
データ	リンク有	20,467	47,965
	リンク無	250,825	22,427,415

閾値: 0.0467, 正答率: 0.987, MCC: 0.145

表 6　混同行列（テストデータ）

		予測	
		リンク有	リンク無
データ	リンク有	1,725	5,659
	リンク無	20,371	2,501,329

閾値: 0.0553, 正答率: 0.990, MCC: 0.131

5　シミュレーション分析

前節では、トピックモデルの推定を行ったが、本節では推定されたパラメータを用いて構築されたネットワークから、インフルエンサーを検出するシミュレーションを行う。シミュレーションを通して、本研究で提案するインフルエンサー検出モデルの有効性を検証することが本節の目的である。ここでのシミュレーションは、Chen et al. (2017) の方法を踏襲し、n 人のユーザーをインフルエンサーとみなし、情報拡散の震源地としたときに、どれほどのユーザーに情報が拡散されるか（＝リーチ率）を検証する。このシミュレーションが想定している状況は、「インフルエンサーがあるトピックに関する投稿を拡散した（投稿と拡散はそれぞれ Twitter 上でのツイートとリツイートに相当する）」という状況であり、二値ネットワークから検出されたインフルエンサーと、重み付きネットワークから検出されたインフルエンサーで、シミュレーション上で両者の情報の拡散力を比較するものである。以下では、シミュレーションの詳細を説明し、結果の解釈を行う。

まず、推定されたパラメータから、リンクごと・トピックごとに重みの異なるネットワークを構築する方法については、3.2.2 項で説明した通りである。具体例として、一部のユーザーを抜き出し、その関係性を二値ネットワークとして表したものを図 7 に、推定パラメータを用いて重み付きネットワークとして表したもの（トピック 6 に関するネットワーク）を図 8 に示す。ただし、矢印の太さは重みの強さを表し、矢印が存在しなければその二者間にはリンクがないことを表している。これを見ると、提案モデルにより構築した重み付きネットワークが、リンクごとトピックごとの関係性の強さを捉えていることや、未観測リンクに対しても潜在的な関係性の強さを反映できていることが分かる。次に、3.2.3 項で説明した方法によって、これら二種類のネットワークに対して次数中心性及び固有ベクトル中心性を計算し、その指標が高いユーザーから順に n 人をインフルエンサーとしてみなす。つまり、各トピックのネットワークに対して、ネットワークの種類・中心性の指標で計 4 種類のインフルエンサー検出手順があり、それぞれの手順で定めたインフルエンサーの拡散力を比較する。そして、リーチ率の計算方法であるが、インフルエンサーとみなされたユーザーとのリンクのうち、4.3 項で求めた閾値である 0.0467 を超えた重みをもったリンク上で情報が拡散され、閾値を超えないリンク上では情報が寸断される、あ

るいは情報が届いても関心を持ってもらえないと仮定することで、全体の何%のユーザーが情報を受け取ったかをリーチ率として定義している。

各ネットワーク及び中心性によって計算された値の上位 n 人をインフルエンサーとし、n を 1 からユーザー総数である 5,028 まで変化させた時のリーチ率を図 9 に示す（紙面の都合上トピック 4・6・9 の結果のみを抜粋している）。次数中心性によって検出したインフルエンサーについては、重み付きネットワークよりも二値ネットワークを仮定して検出した方が、より拡散力を持ったユーザーを見つけ出すことができるという結果であった。一方、固有ベクトル中心性によって検出したインフルエンサーについては、重み付きネットワークを仮定した方がより早く情報が拡散されている。なお、これらの結果は、他のトピックについても概ね同様の傾向を示している。つまり、「影響力の強いリンクを多く集めるユーザー」をインフルエンサーとみなす状況においては、二値ネットワークを、「影響力を持ったユーザーからのリンクを集めるユーザー」をインフルエンサーとみなす状況においては、重み付きネットワークを仮定してインフルエンサーを検出すべきであるとの示唆が得られたことになる。ただし、これらの結果はあくまでシミュレーション上での結果であり、実際の情報拡散過程を追跡したデータを用いた検証ではないことに注意されたい。実データを用いたインフルエンサー検出モデルの検証は今後の課題としたい。

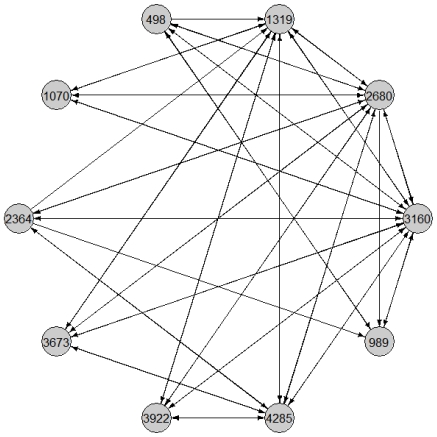

図 7　二値ネットワーク

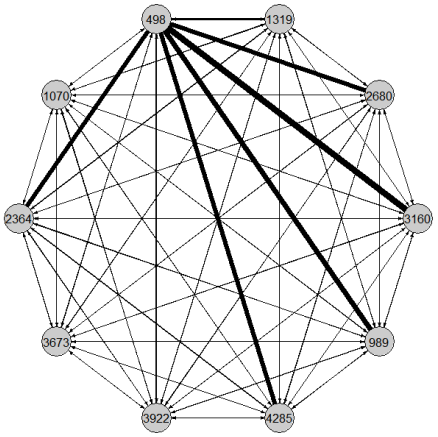

図 8　重み付きネットワーク（トピック 6）

6　結論と今後の課題

SNS の広まりによって世界中の人々の間で社会ネットワークが構築されている現代において、社会ネットワークをモデル化する研究は、企業のマーケティング活動にとって重要なものとなっている。そこで本研究では、社会ネットワーク上のリンク情報だけでなく、ソーシャルメディア上に投稿されるテキスト情報も考慮して社会ネットワーク分析を行うことができるトピックモデルを提案した。さらに、この提案モデルから得られるパラメータを用いて、リンクごと・トピックごとの関係性の強さで重みづけしたネットワークが構築できることを示した。

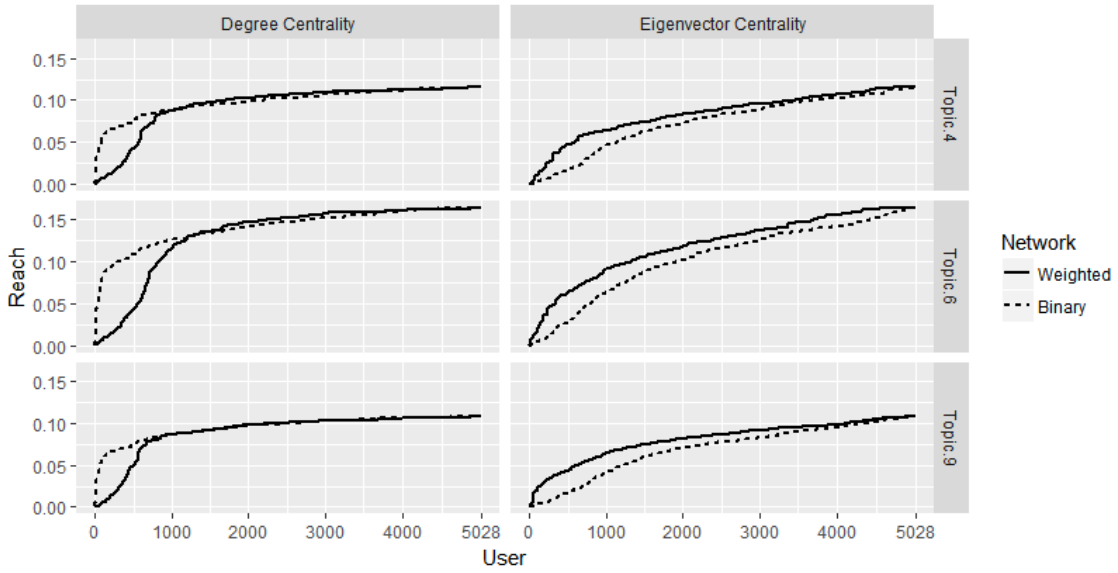

図 9　シミュレーション結果

先行研究で提案されているネットワーク分析モデルとの差異は次の 3 点である。一つ目は、未観測リンクに対する扱いであり、先行研究が未観測リンクに対して一律に 0 と重みづけをしていたのに対し、提案モデルでは、ユーザー間の関係性に潜在的なトピックを仮定することで、未観測リンクであっても潜在的な重みを与えることができる。二つ目は、似ていない他者とのリンクの可能性を考慮することであり、先行研究では、homophily の考え方に基づいて類似度の低いユーザー間のリンクに対して一律に弱い重みを付していた。しかし、ソーシャルメディアが発達した現代では、たとえ共通の趣味や境遇がなかったとしてもリンクが結ばれる可能性を考慮すべきであり、本研究では、トピック間のリンク確率に従って一律でない重みを与えるモデルを提案している。最後は、先行研究でほとんど取り組まれていないソーシャルメディア上のテキスト情報を考慮したモデリングであり、提案モデルは、リンクトピックと単語トピックを相互に対応させてトピック推定を行うことで、テキスト情報の考慮を実現している。この三点の課題を解決することで、より現実に即したネットワークモデルからインフルエンサーの検出を行うことが出来る。

実証分析では、Twitter データを分析して重み付きネットワークを構築し、そこから検出されるインフルエンサーの拡散力を検証するシミュレーションを行った。「影響力を持ったユーザーからのリンクを集めるユーザー」をインフルエンサーをとみなす固有ベクトル中心性の指標においては、二値ネットワークよりも本研究で提案する重み付きネットワークを考慮した方が、より拡散力の高いインフルエンサーを検出することができるという示唆を、シミュレーションの結果から得ることが出来た。しかし、これはあくまでシミュレーションが示す結果であり、実際の社会ネットワーク上で観測された拡散データに対して、モデルの当てはまりのよさを検証するという課題が残されている。Chen et al. (2017) では、複数のネットワークが情報の拡散に与える影響を実データから推定し、拡散過程のモデリングに取り組んでいる。今後は、本研究で提案したネットワーク分析モデルを応用し、拡散過程を説明するモデルに組み込むことでモデルの精緻化を図るなど、提案モデルの応用範囲を広げていきたい。

Appendix

A.1 提案モデルの生成過程と推定アルゴリズム

2 節で示した提案モデルの詳細な生成過程は以下の通りである。

1. For トピック $k = 1, \ldots, K$
   (a) 単語分布を生成　　$\phi_k \sim Dirichlet(\beta)$
   (b) For トピック $k' = 1, \ldots, K$
      (i) For ユーザー $d = 1, \ldots, D$
         リンク確率を生成　　$\psi_{kk'}^{(d)} \sim Beta(\gamma, \delta)$
2. For ユーザー $d = 1, \ldots, D$
   (a) トピック分布を生成　　$\theta_d \sim Dirichlet(\alpha)$
   (b) For ユーザー $d' = 1, \ldots, D$
      (i) リンクトピック (sender) を生成　　$S_{dd'} \sim Categorical(\theta_d)$
      (ii) リンクトピック (receiver) を生成　　$R_{dd'} \sim Categorical(\theta'_d)$
      (iii) リンクを生成　　$y_{dd'} \sim Bernoulli(\psi_{S_{dd'}R_{dd'}}^{(d)})$
   (c) For 単語 $n = 1, \ldots, M_d$
      (i) 単語トピックを生成　　$z_{dn} \sim Categorical(\frac{N_{d1}}{2(D-1)}, \cdots, \frac{N_{dK}}{2(D-1)})$
      (ii) 単語を生成　　$w_{dn} \sim Categorical(\phi_{z_{dn}})$

提案モデルの推定には崩壊型ギブスサンプリングが用いられており、リンクトピック及び単語トピックのサンプリング式は以下の通りである。

$$
\begin{aligned}
p(S_{dd'} = k, R_{dd'} = k' | S_{dd'}, R_{dd'}, Y, Z, \alpha, \gamma, \delta) \propto \\
&\left(N_{dk\backslash dd'} + \alpha_k\right) \times \left(N_{d'k'\backslash dd'} + \alpha'_k\right) \times \\
&\left(\frac{N_{dk\backslash dd'} + 1}{N_{dk\backslash dd'}}\right)^{M_{dk}} \times \left(\frac{N_{d'k'\backslash dd'} + 1}{N_{d'k'\backslash dd'}}\right)^{M_{d'k'}} \times \\
&\frac{\left(n_{kk'\backslash dd'}^{(+,d')} + \gamma_{kk'}\right)^{\mathbb{I}(y_{dd'}=1)} \left(n_{kk'\backslash dd'}^{(-,d')} + \delta_{kk'}\right)^{\mathbb{I}(y_{dd'}=0)}}{n_{kk'\backslash dd'}^{(+,d')} + n_{kk'\backslash dd'}^{(-,d')} + \gamma_{kk'} + \delta_{kk'}}
\end{aligned}
\tag{11}
$$

$$
p(z_{dn} = k | Z_{\backslash dn}, W, S, R, \beta) \propto N_{dk} \times \left(\frac{M_{kw_{dn}\backslash dn} + \beta_{w_{dn}}}{\sum_{v=1}^{V} M_{kv\backslash dn} + \beta_v}\right) \tag{12}
$$

ただし、N_{dk}, M_{dk}, M_{kv} は、それぞれユーザー d が関係するリンクのうちトピック k が割り当てられた数、ユーザー d の単語のうちトピック k が割り当てられた数、語彙 v にトピック k が割り当てられた回数を表しており、$n_{kk'}^{(+,d)}(n_{kk'}^{(-,d)})$ は、トピック k から k' のリンクのうちユーザー d が関係していて、かつリンクが結ばれている（結ばれていない）数を表している。また、$\mathbb{I}$ は、括弧内の条件が満たされたときに 1 を、満たされなかったときに 0 を返す関数であり、$\backslash dd'(\backslash dn)$ 記号は、該当するリンク（単語）をカウントから除く操作を表している。

このようにして得られたトピック割り当てのカウント数を用いて、各パラメータの推定値は次のよう

にして得られる。

$$\theta_{dk} = \frac{N_{dk} + \alpha_k}{\sum_{k=1}^{K} (N_{dk} + \alpha_k)} \tag{13}$$

$$\psi_{kk}^{(d)} = \frac{n_{kk'}^{(+,d)} + \gamma_{kk'}}{n_{kk'}^{(+,d)} + n_{kk'}^{(-,d)} + \gamma_{kk'} + \delta_{kk'}} \tag{14}$$

$$\phi_{kv} = \frac{M_{kv} + \beta_v}{\sum_{v=1}^{V} (M_{kv} + \beta_v)} \tag{15}$$

4 節ではこのサンプリング式に従ってサンプリングが行われているが、その時のハイパーパラメータの設定は表 7 の通りである。

表 7　ハイパーパラメータの設定

ハイパーパラメータ	α	β	γ	δ
設定	$1/K$ (K はトピック数)	$1/V$ (V は語彙数)	2.0　(対角成分) 1.0　(非対角成分)	1.0　(対角成分) 2.0　(非対角成分)

A.2 Perplexity の定義式

y^{test}, w^{test} をテストデータセット内のリンク及び単語とすると、テストデータに対する Perplexity を計算するためには、予測確率 $p(y^{test}, w^{test}|model)$ を計算する必要がある。しかし、これは解析的に求めることができないため、ギブスサンプリングによる各サンプル s によって以下のように近似する。

$$p(y^{test}, w^{test}|model) = p(y^{test}|model)p(w^{test}|model) \tag{16}$$

$$p(y_{dd'}^{test}|model) \sim \frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{k'=1}^{K} \frac{N_{dk}^{(s)} + \alpha_k}{\sum_{l=1}^{K}\left(N_{dl}^{(s)} + \alpha_l\right)} \frac{N_{d'k'}^{(s)} + \alpha_{k'}}{\sum_{l=1}^{K}\left(N_{d'l}^{(s)} + \alpha_l\right)} \frac{\left(n_{kk'\backslash dd'}^{(+,d',s)} + \gamma_{kk'}\right)^{\mathbb{I}(y_{dd'}^{test}=1)} \left(n_{kk'\backslash dd'}^{(-,d',s)} + \delta_{kk'}\right)^{\mathbb{I}(y_{dd'}^{test}=0)}}{n_{kk'\backslash dd'}^{(+,d',s)} + n_{kk'\backslash dd'}^{(-,d',s)} + \gamma_{kk'} + \delta_{kk'}} \tag{17}$$

$$p(w_{dn}^{test}|model) \sim \frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K} \frac{N_{dk}^{(s)}}{\sum_{l=1}^{K} N_{dl}^{(s)}} \frac{M_{kw_{dn}^{test}} + \beta_{w_{dn}^{test}}}{\sum_{u=1}^{V}\left(M_{ku}^{(s)} + \beta_u\right)} \tag{18}$$

この近似予測確率を用いて Perplexity は次のように計算される。

$$Perplexity = exp\left(\frac{\sum_{d=1}^{D}\sum_{d'=1}^{D}\log p(y_{dd'}^{test}|model)}{\sum_{d=1}^{D} N_d^{test}} + \frac{\sum_{d=1}^{D}\sum_{n=1}^{M_d}\log p(w_{dn}^{test}|model)}{\sum_{d=1}^{D} M_d^{test}}\right) \tag{19}$$

ただし、N_d^{test} 及び M_d^{test} は、テストデータセット内のユーザー d に関係するリンクの数、ユーザー d の単語数を表す。

A.3 テキストデータの前処理

取得したテキストデータに対して、以下の順に前処理を行った。

1. ストップワードの除去

統計ソフト R 内の「tm」パッケージに含まれている英語のストップワードに加えて、記号・URL 文字列・スクリーンネーム（@+ 英数字の文字列）などをストップワードとしてデータから取り除いた。

2. ステミング（テキストの正規化）

全ての単語について小文字に統一し、ステミングの処理を行った。ステミングとは、単語の活用形を直す正規化処理のことで、例えば、COME、came、coming などの単語は全て come に統一される。なお、このステミングのアルゴリズムには、「tm」パッケージ内の関数である「stemDocument」が用いられている。

3. 形態素解析による名詞の抽出

単語の品詞を検出する形態素解析を行い、名詞のみを使用するデータとして残した。なお、形態素解析のアルゴリズムには、「TreeTagger」というソフトウェアが用いられている。

4. TF-IDF による単語の選定

トピックの解釈性向上と計算コストの削減を目的に、TF-IDF と呼ばれる指標を用いて単語の重要度を推定し、上位 1% の単語のみを使用するデータとして残した。TF-IDF は、以下の式で計算される。

$$TFIDF_{dv} = \frac{N_{dv}}{\sum_{v=1}^{V} N_{dv}} \times \log\left(\frac{D}{D_v} + 1\right) \tag{20}$$

ただし、D_v は、語彙 v が出現する文書数を表す。

注釈

1. これらのトピックモデルの詳細については、佐藤（2015）と岩田（2015）などが参考になる。
2. 中心性の指標を含む社会ネットワーク分析の諸手法については、例えば金光（2003）などが参考になる。
3. 2018 年 7 月 27 日時点で 9,388,292 人のフォロワー数であった。
4. 日本時間 2018 年 3 月 9 日に「Nintendo Direct」と呼ばれる新作ゲームソフトの発表イベントがあり、ユーザーが一斉にこのイベントに関する投稿を行っていることが確認された。そのため本研究では、これら全ユーザーで共通するような情報がトピックの推定に与える影響を避けることを目的として、投稿データの取得期間を 2018 年 2 月 28 日までとしている。
5. 一般的なギブスサンプリングでは、各繰り返しで得られたサンプルを平均することでパラメータの推定値を得るが、崩壊型ギブスサンプリングでは、サンプリングしたトピック割り当てのカウントデータを推定に用いるため、最後の繰り返しで得られたトピックを使用しても問題ない。岩田（2015）では、トピックモデルの推定に関して詳細にアルゴリズムを説明しており、参考になる。
6. ほとんどリンクがないユーザー 1 の Topic 1 のや、列のリンク確率が高い値で推定されているのは、事前分布がそのまま反映されているためである（Appendix を参照してほしい）。
7. 表内の単語は、前処理で行ったステミングの結果を載せているため、「#gamedevelopment」や「nowplaying」といった単語の語幹となっていることに注意されたい。

参考文献

[1] Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9(Sep), 1981-2014.

[2] Ansari, A., Koenigsberg, O., and Stahl, F. (2011), "Modeling Multiple Relationships in Social Networks," *Journal of Marketing Research*, 48(4), 713-728.

[3] Aral, S. and Walker, D. (2014), "Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment," *Management Science*, 60(6), 1352-1370.

[4] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3(Jan), 993-1022.

[5] Blei, D. M. and Jordan, M. I. (2003), "Modeling Annotated Data," In *Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 127-134.

[6] Borgatti, S. P. (2005), "Centrality and Network Flow," *Social Networks*, 23(3), 191-201.

[7] Braun, M. and Bonfrer, A. (2011), "Scalable Inference of Customer Similarities from Interactions Data Using Dirichlet Processes," *Marketing Science*, 30(3), 513-531.

[8] Buckley, F. and Harary, F. (1990), *Distance in Graphs*. Reading, MA: Addison-Wesley.

[9] Chen, X., van der Lans, R., and Phan, T. Q. (2017), "Uncovering the Importance of Relationship Characteristics in Social Networks: Implications for Seeding Strategies," *Journal of Marketing Research*, 54(2), 187-201.

[10] Freeman, L. C. (1979), "Centrality in Social Networks I: Conceptual Clarification," *Social Networks*, 1, 215-239.

[11] Godes, D. and Mayzlin, D. (2009), "Firm-Created Word-of-Mouth Communication: Evidence from a Field Test," *Marketing Science*, 28(4), 721-739.

[12] Granovetter, M. (1973), "The Strength of Weak Ties," *American Journal of Sociology*, 78(6), 1360-1380.

[13] Hinz, O., Skiera, B., Barrot, C., and Becker, J. U. (2011), "Seeding Strategies for Viral Marketing: An Empirical Comparison," *Journal of Marketing*, 75(6), 55-71.

[14] Hu, Y. and Van den Bulte, C. (2014), "Nonmonotonic Status Effects in New Product Adoption," *Marketing Science*, 33(4), 509-533.

[15] Iyengar, R., Van den Bulte, C., and Valente, T. W. (2011), "Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30(2), 195-212.

[16] Matthews, B. W. (1975), "Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2), 442-451.

[17] Moschis, G. P., and Moore, R. L. (1979), "Decision Making among the Young: A Socialization Perspective," *Journal of Consumer Research*, 6(2), 101-112.

[18] Park, E., Rishika, R., Janakiraman, R., Houston, M. B., and Yoo, B. (2018), "Social Dollars in Online Communities: The Effect of Product, User, and Network Characteristics," *Journal of Marketing*, 82(1), 93-114.

[19] Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009),“Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora,” In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1(1), 248-256.

[20] Reingen, P. H., Foster, B. L., Brown, J. J., and Seidman, S. B. (1984),“Brand Congruence in Interpersonal Relations: A Social Network Analysis,” *Journal of Consumer Research*, 11(3), 771-783.

[21] Sweet, T. M. and Zheng, Q. (2018), “Estimating the Effects of Network Covariates on Subgroup Insularity with a Hierarchical Mixed Membership Stochastic Blockmodel,” *Social Networks*, 52, 100-114.

[22] Trusov, M., Bodapati, A. V., and Bucklin, R. E. (2010), “Determining Influential Users in Internet Social Networks,” *Journal of Marketing Research*, 47(4), 643-658.

[23] Van den Bult, C., and Wuyts, S. (2008), *Social Networks and Marketing*. Cambridge, MA: Marketing Sience Institute.

[24] 岩田具治（2015）『トピックモデル』, 機械学習プロフェッショナルシリーズ, 講談社.

[25] 佐藤一誠（2015）『トピックモデルによる統計的潜在意味解析』, 自然言語処理シリーズ, コロナ社.

[26] 金光淳（2003）『社会ネットワーク分析の基礎：社会的関係資本論に向けて』, 勁草書房.

[27]“第 1 部　第 4 章　ICT によるインクルージョン,” 情報通信白書, 平成 30 年度, 総務省, 151-205, http://www.soumu.go.jp/johotsusintokei/whitepaper/index.html, (参照 2018-07-16).