統計数理(2021)
第 69 巻 第 1 号 99–118

［研究ノート］

ソーシャルメディア上のテキスト情報を考慮した社会ネットワーク分析モデル

―次数異質性モデルへの拡張―

五十嵐 未来[†]・照井 伸彦[†]

(受付 2020 年 5 月 25 日；改訂 10 月 6 日；採択 11 月 2 日)

要　　旨

近年，社会ネットワークをモデル化して分析する際に，ネットワーク情報だけでなく，人々がソーシャルメディア上などで生成するテキスト情報を考慮してコミュニティ構造を捉えることの重要性が増している．テキスト情報を考慮することにより，ネットワーク上で密にエッジが形成されている構造の中に，人々が持つ興味や関心に応じた複数のまとまりが存在するというような複雑なコミュニティ構造を持つ社会ネットワークの分析が可能となる．本研究では，これをモデル化した先行研究によるネットワークデータとテキストデータの同時利用モデルを拡張し，社会ネットワークにおいて一般的な性質であるエッジの生成されやすさがノードごとに異なる次数異質性を考慮したモデルを提案する．Twitter を用いた実証分析では，テキスト情報の活用及び次数異質性の考慮が予測性能に与える影響を検証するため，複数の比較モデルと共に比較実験を行い，提案モデルが優れた予測性能を持つことを示した．

キーワード：社会ネットワーク分析，コミュニティ検出，テキスト解析，トピックモデリング，ベイズ推定，ノード次数異質性．

1. はじめに

Social Networking Sites(SNS)の流行や電子商取引サイトの台頭などにより，消費者を取り巻く社会ネットワークを分析し，その構造を把握することは，企業のマーケティング活動における重要な位置を占めるようになっている．社会ネットワーク分析の手法は，統計学や社会学の分野を中心に長年研究されており，コミュニティ構造の抽出に代表されるように，ネットワークデータを要約し理解するための統計モデルが多く提案されている(e.g., Snijders and Nowicki, 1997; Airoldi et al., 2008)．これらのモデルでは，ネットワーク上のノードとエッジを観測データとして扱い，他と比べてエッジ密度が高くなるノード集合として定義されるコミュニティ構造を抽出する．また，社会ネットワークにおけるノードは，人々のことを表しており，人々の属性や行動といった付随的なデータを考慮することで，ネットワークモデルの精緻化を目指す研究も熱心に取り組まれている(e.g., Handcock et al., 2007)．中でも，近年は，ソーシャルメディアの流行や口コミ機能を搭載した電子商取引サイトの台頭などにより，ノードごとに固有のデータとしてユーザー生成コンテンツ(User-Generated-Contents，UGC)，特にテキスト情報

[†] 東北大学 経済学研究科：〒 980–8576 宮城県仙台市青葉区川内 27–1

をネットワーク情報と組み合わせた社会ネットワーク分析モデルが多く提案されている(e.g., Liu et al., 2009; Bouveyron et al., 2018).

ネットワーク情報だけでなくテキスト情報も考慮したモデルを構築することの利点は，一方の情報だけでは捉えることが難しいコミュニティ構造を識別できることである．ネットワーク情報のみを考慮する従来の研究では，他と比べてエッジ密度が高くなるノード集合をコミュニティと定義するが，テキスト情報も考慮してネットワークモデルを構築する近年の研究では，エッジの密度だけでなく，トピック比率などを用いたテキスト情報の類似性も考慮してコミュニティを定義している．例えば，Igarashi and Terui (2020)では，そのようなコミュニティをトピックベース・コミュニティと名付け，ネットワークのみ，もしくはテキストのみを利用するモデルと比較して，両者を考慮するモデルの方が精度よくコミュニティの分割が可能となることを示している．より具体的にトピックベース・コミュニティの概念を説明するために，例えば，ある学校の同級生で構成されるコミュニティを想定する．そこでは，学生らは互いに何らかの関係性を持った密度の高いネットワークが形成されているはずである．したがって，ネットワーク情報のみを考慮したモデルを用いると，そのようなネットワーク上には，一つのコミュニティが存在していると認識される．しかし，それと同時に，学生らは音楽や読書，スポーツといった様々な趣味を持っていることが考えられるため，共通の趣味を持った学生らをまとめて複数のコミュニティが存在するとみなす方が，より意味のあるセグメンテーションとなる可能性がある．ソーシャルメディアに代表されるオンライン上の社会ネットワークでは，上で説明したような，現実世界における社会的なつながりの上に存在する，興味や関心などに基づいたつながり，つまりトピックベース・コミュニティが点在していることが考えられる．したがって，ソーシャルメディア上に生成されたテキストコンテンツからそのユーザーの興味や関心を推定してネットワーク情報と結びつけることは，社会ネットワーク分析モデルを精緻化させることに繋がるだけでなく，複雑多様な現代のオンライン社会ネットワークの構造を理解するうえで有意義な分析となりうる．

また，社会ネットワークが持つ性質の一つとして，次数がノードごとに異質的であるという性質(次数異質性)がある．これは，多くの人々が少数の人とだけネットワーク上で関係性を持つ一方で，限られた人々が多くの人々と関係性を持つ傾向にあるという性質である．このように，現実の社会ネットワークにおいて，エッジが結ばれる確率は一定ではなく，ノードごとに異質であると仮定する方が，より現実に即したモデリングと言えるが，確率的ブロックモデル(Snijders and Nowicki, 1997; Igarashi and Terui, 2020)など代表的なネットワークモデルの多くでは，次数異質性を考慮していない．本研究では，ブロックモデルに即したエッジ生成確率をノードごとに異なるパラメータとして推定することで Igarashi and Terui (2020)のモデルを拡張し，次数異質性を考慮したモデルを提案する．実証分析では，現実のソーシャルメディアとして Twitter から得られたデータセットを用い，提案モデルである次数異質性を考慮した上でネットワーク情報とテキスト情報を結びつける社会ネットワークモデルを推定する．また，提案モデルに含まれる次数異質性及びテキスト情報の利用が外挿予測に与える効果を検証するために，提案モデルからそれぞれの特徴を除いた比較モデルとの比較実験を行う．

以下，2 節では，社会ネットワーク分析に関係する先行研究をまとめ，本研究の目的と位置づけを明確にする．3 節では，提案モデルを説明し，4 節ではその推定法を導出する．続いて，5 節では，Twitter データを利用した実証研究を報告し，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 社会ネットワーク分析モデルの進展

統計学や社会学などを中心として，古くから社会ネットワークをモデル化し，その構造を把握するための研究が続いている．中でも代表的なものが，確率的ブロックモデル（Stochastic Block Models, SBM, Wang and Wong, 1987; Snijders and Nowicki, 1997）である．SBM は，ノードが K 個のコミュニティのうち一つだけに属することを仮定しており，ノード i が属するコミュニティを $z_i \in \{1, \ldots, K\}$ とすると，ノード i と j の間にエッジが生成される確率は，$\psi_{z_i z_j}$ で表される．これは，$K \times K$ 行列 Ψ の (z_i, z_j) 成分であり，エッジ確率を表すパラメータである．

SBM は，様々な文脈でモデルの拡張が行われている．SBM がノードに単一のメンバーシップを仮定していたのに対し，Airoldi et al.（2008）は，各ノードが他ノードとの関係性ごとに異なるコミュニティに属することを許容する混合メンバーシップ確率的ブロックモデル（Mixed Membership Stochastic Blockmoldels, MMSB）を提案している．ノード i から j の関係性において，ノード i が属するコミュニティを s_{ij}（sender），ノード j が属するコミュニティを r_{ji}（receiver）とすると，両者の間にエッジが生成される確率は，$\psi_{s_{ij} r_{ji}}$ で表される．この拡張により，MMSB はコミュニティの重なりを考慮することができ（SBM ではコミュニティが重なることはない），より現実に即したモデリングが可能となっている．

また，社会学の文脈では，ノード間の関係性が性別や年齢といったノード固有の特徴量の影響を受けて決まることも知られている（Hoff et al., 2002; Handcock et al., 2007; Krivitsky et al., 2009）．しかし，本研究では，ソーシャルメディアに代表されるようなオンライン上の社会ネットワークに着目しているため，そのような特徴量は考慮しない．Twitter のような匿名型ソーシャルメディアでは，ユーザーは年齢や性別といった個人情報を隠した状態でアカウントを登録することができ，そのような状況において他者と関係を結ぶ際に考慮できる情報は，相手が形成しているネットワークとメディア上に投稿したコンテンツのみである．ただし，そのようなノードの属性を示すデータが利用可能であれば，提案モデルに取り込むことは容易であり，社会学的視点からの分析も可能である．

2.2 ネットワークとテキスト情報の同時モデリングに関する研究

前節で挙げた社会ネットワークモデルに関する研究では，ネットワーク情報のみに着目してモデルを提案しているが，近年，Twitter や Facebook といったオンライン上の社会ネットワーク構造をより深く理解するために，ネットワークとテキスト情報をどちらも考慮するモデルが盛んに研究されている．例えば，Chang and Blei（2010）は，ノードに固有のテキスト情報に対してトピックモデルを適用し，ノードのテキストに割り当てられたトピック割合の類似度に応じてノード間のエッジ生成確率が定義される関係トピックモデル（Relational Topic Model, RTM）を提案している．ただし，RTM の目的が，ネットワーク情報を加味してテキスト情報におけるトピックを推定するのに対して，Igarashi and Terui（2020）及び本研究のモデルは，テキスト情報を考慮してネットワーク上のコミュニティ構造を把握する点で対照的である．

Chang and Blei（2010）のようにテキスト情報を潜在的ディリクレ配分法（latent Dirichlet allocation,LDA, Blei et al., 2003）やその拡張モデルを用いてネットワークモデルに取り込むという方法は他にもいくつかの研究で見られる．例えば，Liu et al.（2009）は，Topic-Link LDA を提案しており，ノード固有のテキスト情報を考慮してコミュニティ構造を検出するという点で本研究と同じ目的を持っている．ただし，SBM と同様に，ノードが単一のコミュニティに属することを仮定した限定的な研究である．また，Liu et al.（2009）では，エッジ生成確率

表 1. 提案モデルと既存モデルの比較.

	観測データ	メンバーシップ	グラフの方向性	次数異質性
Blei et al. (2003)	テキストのみ	混合	-	-
Snijders and Nowicki (1997)	ネットワークのみ	単一	両方可能	考慮せず
Airoldi et al. (2008)	ネットワークのみ	混合	両方可能	考慮せず
Chang and Blei (2010)	ネットワーク／テキスト	混合	無向グラフのみ	考慮せず
Liu et al. (2009)	ネットワーク／テキスト	単一	無向グラフのみ	考慮せず
Bouveyron et al. (2018)	ネットワーク／テキスト	単一	両方可能	考慮せず
Zhu et al. (2013)	ネットワーク／テキスト	混合	両方可能	考慮せず
Igarashi and Terui (2020)	ネットワーク／テキスト	混合	両方可能	考慮せず
Karrer and Newman (2011)	ネットワークのみ	単一	両方可能	ノードごとの期待次数パラメータを導入
本研究	ネットワーク／テキスト	混合	両方可能	エッジ確率を異質パラメータとして定義

が，ノード固有のトピック及びコミュニティ割合の類似度によって定義されているため，エッジの向きが逆になってもその生成確率が変わらない，つまり無向グラフを想定しているのに対し，本研究を含めたブロックモデルにおいては，$K \times K$ 行列のエッジ確率パラメータを用いたネットワークモデリングにより，グラフの方向性にかかわらずモデルを適用可能である．他にも，Bouveyron et al. (2018)は，SBM にテキスト情報のモデルを加える形で拡張した Stochastic Topic Block Model(STBM)を提案している．

これらは単一のメンバーシップを仮定した SBM の拡張モデルであるが，Zhu et al. (2013)は，ノードの混合メンバーシップを仮定し，テキストとネットワーク情報の両者を考慮するネットワーク分析モデルを提案している．本研究における提案モデルとの相違点は，Zhu et al. (2013)は，エッジに割り当てられるコミュニティと単語に割り当てられるトピックが同一の分布に従っているという点であり，言い換えれば，コミュニティとトピックの次元を同一のものとして扱っている．しかし，現実の社会ネットワークでは，コミュニティとトピックが必ずしも互いに対応しているとは限らない．例えば，音楽とスポーツに興味のある人々が同じコミュニティ内に存在するネットワークを考える．このようなコミュニティを Zhu et al. (2013)のモデルで検出したとすると，一つのコミュニティに対して，音楽とスポーツという複数の意味的まとまりをもつトピックが対応してしまい，トピックの解釈性に欠ける．一方で，Igarashi and Terui (2020)及び本研究では，コミュニティとトピックがそれぞれ異なる分布に従うことを仮定しており，上記のようなネットワークに対しても，一つのコミュニティと，音楽トピック及びスポーツトピックのように別々に複数トピックを対応させることができる．3 節では，その詳細な定式化を説明する．

これらの既存モデルを踏まえて，本研究では，Igarashi and Terui (2020)によるノードの混合メンバーシップを仮定したネットワークとテキストの同時モデリングを拡張し，エッジ確率をノードごとに異質なパラメータとするモデルを検討する．これにより，社会ネットワークが一般的に有する次数異質性を考慮したモデリングが可能となる．エッジ生成確率の異質性については，Karrer and Newman (2011)が，ノードごとの期待次数をパラメータとして導入し，関係するノードに応じてエッジ生成確率が異質となるような補正を行うモデルを提案している．一方，本研究は，エッジ生成確率自体をノードごとに異質なパラメータとして直接推定する．

表 1 では，ここまでに議論した本研究と先行研究との比較をまとめている．まず，ネットワークやテキストどちらかのみを観測データとして扱うモデルと比較すると，本研究で提案するモデルは，その両者を考慮して社会ネットワーク分析を行うものであり，前述したようにどちらか一方の情報だけでは捕捉することが難しいネットワーク構造を明らかに出来る可能性がある．また，その両情報を扱う既存モデルと比較すると，ノードに混合メンバーシップを許容

している点，グラフの有向無向にかかわらず適用可能な点，そして社会ネットワークにおける次数異質性を考慮したモデリングを行っている点が本研究の特徴である．

3. モデル

本節では，まず提案モデルの基礎となる Igarashi and Terui (2020)のモデルを説明し，次にその差異を明らかにしながら本研究で使用するモデルの説明を行う．また，両モデルで共通して，観測されるデータは，ネットワーク情報を表す隣接行列 A，及びノードに固有のテキスト情報を表す単語の Bag-of-Words 集合 W の二つである．

まず，D 個のノードを持つ有向グラフを考えると，その隣接行列 A は，$D \times D$ 行列であり，行列の各要素はノード間の関係性を示す二値変数である．つまり，$a_{ij} = 0$ はエッジが存在しないことを表し，$a_{ij} = 1$ は存在することを表す．また，自己ループは考えないこととし，全ての i について $a_{ii} = 0$ である．Igarashi and Terui (2020)では，ノード i から j への関係性において，その送り手 i が潜在的なコミュニティ $s_{ij} \in \{1, \ldots, K\}$($K$ はコミュニティ数)に属し，受け手 j が潜在コミュニティ $r_{ji} \in \{1, \ldots, K\}$ に属することを仮定する．また，これら潜在コミュニティの行列表現を $S = (s_{ij}), R = (r_{ji})$ とする．モデルの生成過程において，送り手及び受け手のコミュニティはカテゴリカル分布，$s_{ij} \mid \eta_i \sim Categorical\,(\eta_i)$，$r_{ji} \mid \eta_j \sim Categorical\,(\eta_j)$ に従う．ただし，$\eta_i = (\eta_{i1}, \ldots, \eta_{iK})^\top$ はノード i のコミュニティ所属割合を表すパラメータであり，$\sum_k \eta_{ik} = 1$ を満たす．このコミュニティ分布の行列表現は $H = (\eta_1, \ldots, \eta_D)$ で表される．H は事前分布としてディリクレ分布 $\eta_i \mid \gamma \sim Dirichlet(\gamma)$ に従うことを仮定しており，$\gamma = (\gamma_1, \ldots, \gamma_K)^\top$ は推定にあたって調整が必要なハイパーパラメータである．

ノード i と j 間の関係性 a_{ij} は，s_{ij} と r_{ji} が所与の時，ベルヌーイ分布，$a_{ij} \mid s_{ij} = k, r_{ji} = k', \psi_{kk'} \sim Bernoulli\,(\psi_{kk'})$ に従うことを仮定する．ただし，$\psi_{kk'}$ は，送り手のコミュニティが k，受け手のコミュニティが k' の時にエッジが生成される確率を示す．また，エッジ確率の $K \times K$ 行列表現は $\Psi = (\psi_{kk'})$ で表され，行列の各要素は，事前分布としてベータ分布，$\psi_{kk'} \mid \delta_{kk'}, \epsilon_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'})$ に従う．このとき，δ, ϵ は Ψ と同じ次元を持つハイパーパラメータである．

従って，コミュニティ分布 H を所与としたときのネットワークデータに対する条件付尤度は以下で定義される．

$$
\begin{aligned}
(3.1)\ & p(A, S, R, \Psi \mid H) \\
&= p(A \mid S, R, \Psi) p(S \mid H) p(R \mid H) p(\Psi \mid \delta, \epsilon) \\
&= \prod_{i=1}^{D} \left\{ \prod_{j=1, j \neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \right\} \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{kk'} \mid \delta_{kk'}, \epsilon_{kk'}).
\end{aligned}
$$

続いて，ノード固有のテキスト情報について考える．ここでは，ノード i が生成したテキストについて，文章内の単語の順番を無視して，つまり Bag-of-Words の形式で保存した M_i 個の単語を観測データとする．ノード i に関する m 番目の単語 w_{im} は潜在的なコミュニティ $x_{im} \in \{1, \ldots, K\}$ 及びトピック $z_{im} \in \{1, \ldots, L\}$($L$ はトピック数)を持つことを仮定する．単語コミュニティと単語トピックの配列表現はそれぞれ X と Z で表され，各配列の要素は M_i 次元のベクトルである．モデルの生成過程において，単語コミュニティ x_{im} はカテゴリカル分布 $x_{im} \mid \eta_i \sim Categorical(\eta_i)$ に従う．ここで，η_i が単語コミュニティ x_{im} だけでなく，ノードコミュニティ s_{ij}, r_{ji} を生成するパラメータであったことを思い出すと，η_i はネットワークデータとテキストデータのモデルに共通するパラメータであり，両者の情報をつなげる役割を果たしている．一方，単語トピックは単語コミュニティが所与の状態でカテゴリカル分布

$z_{im} \mid x_{im} = k, \Theta \sim Categorical(\theta_k)$ に従う．このとき，$\theta_k = (\theta_{k1}, \ldots, \theta_{kL})^\top$ は，コミュニティ k に関するトピック割合を示すパラメータであり，$\sum_l \theta_{kl} = 1$ を満たす．このトピック分布の行列表現は $\Theta = (\theta_1, \ldots, \theta_K)$ であり，事前分布はディリクレ分布 $\theta_k \mid \alpha \sim Dirichlet(\alpha)$ に従う．

単語トピック z_{im} を所与として，それに対応する単語 $w_{im} \in \{1, \ldots, V\}$（$V$ は総単語数）は，単語トピックに対応するカテゴリカル分布 $w_{im} \mid z_{im} = l, \Phi \sim Categorical(\phi_l)$ に従う．ただし，$\phi_l = (\phi_{l1}, \ldots, \phi_{lV})^\top$ は，そのトピックにおいて単語が生成される確率を表す単語分布であり，$\sum_v \phi_{lv} = 1$ を満たす．単語分布の行列表現は $\Phi = (\phi_1, \ldots, \phi_L)$ であり，その事前分布はディリクレ分布 $\phi_l \sim Dirichlet(\beta)$ に従う．

従って，テキストデータに対する条件付尤度は，同じくコミュニティ分布 H を所与として，以下で定義される．

(3.2) $p(W, X, Z, \Theta, \Phi \mid H)$

$$
\begin{aligned}
&= p(W \mid Z, \Phi) p(Z \mid X, \Theta) p(X \mid H) p(\Theta \mid \alpha) p(\Phi \mid \beta) \\
&= \prod_{i=1}^{D} \left\{ \prod_{m=1}^{M_i} \{ p(w_{im} \mid z_{im}, \Phi) p(z_{im} \mid x_{im}, \Theta) p(x_{im} \mid \eta_i) \} \right\} \prod_{k=1}^{K} p(\theta_k \mid \alpha) \prod_{l=1}^{L} p(\phi_l \mid \beta).
\end{aligned}
$$

コミュニティ分布 H を所与とすることで，式(3.1)及び(3.2)の条件付尤度が独立となる仮定を置いているため，Igarashi and Terui (2020)の結合分布は，式(3.1)と(3.2)及び H の密度を掛け合わせることで以下のように得られる．

(3.3) $p(A, W, S, R, X, Z, H, \Psi, \Theta, \Phi)$

$$
\begin{aligned}
&= \prod_{i=1}^{D} \left\{ \prod_{j=1, j \neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \right. \\
&\qquad \left. \prod_{m=1}^{M_i} \{ p(w_{im} \mid z_{im}, \Phi) P(z_{im} \mid x_{im}, \Theta) p(x_{im} \mid \eta_i) \} \right\} \times \\
&\prod_{i=1}^{D} p(\eta_i \mid \gamma) \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{kk'} \mid \delta_{kk'}, \epsilon_{kk'}) \prod_{k=1}^{K} p(\theta_k \mid \alpha) \prod_{l=1}^{L} p(\phi_l \mid \beta).
\end{aligned}
$$

Igarashi and Terui (2020)のモデルでは，ユーザーが生成したテキストコンテンツを考慮しながらネットワーク上のコミュニティ構造を把握する，つまりトピックベース・コミュニティを見つけることを目的としている．このとき，ノード間にエッジが生成される確率を，$a_{ij} = 1 \mid s_{ij} = k, r_{ji} = k' \sim Bernoulli(\psi_{kk'})$ として全てのノードに対して同質的であることを仮定している．しかし，前節でも説明したように，現実の社会ネットワークにおいては，次数がノードによって大きく異なることが一般的であり，Igarashi and Terui (2020)では，この性質を考慮できていないため，現実のネットワークデータに対して十分に適合できない可能性がある．

本研究では，この問題を解決するために，エッジ生成確率の部分を $a_{ij} \mid s_{ij} = k, r_{ji} = k' \sim Bernoulli(\psi_{jkk'})$ としてモデルを拡張する．このとき，$\psi_{jkk'}$ は，送り手のコミュニティが k で，受け手のコミュニティが k' の時にエッジが生成される確率を示し，受け手のノード j に依存する異質なパラメータである．この定式化により，例えば，受け手 j がコミュニティ k の中で多くのエッジを集める，いわゆるハブノードである場合に，$\psi_{jkk'}$ が大きな値を取ることでそれを表現する．これにより，提案モデルは，社会ネットワークにおける次数分布の異質性を反映し，ノードごとの次数の多寡に応じてエッジ確率パラメータを異質的に推定することで，

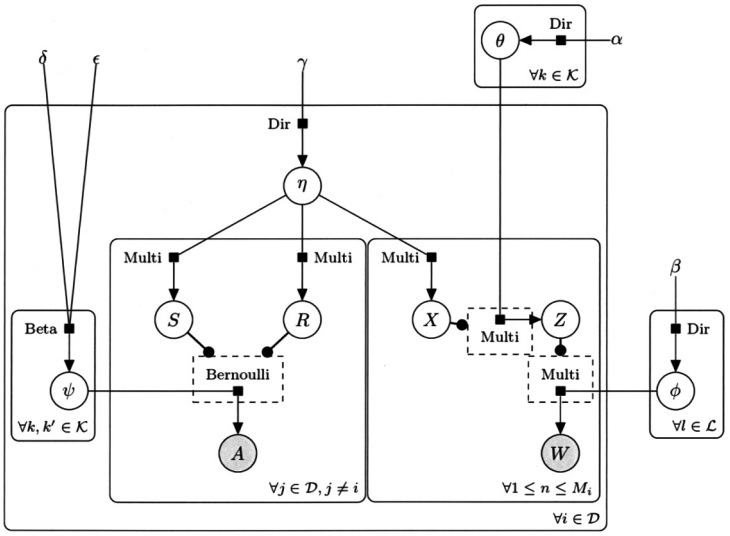

図 1. 提案モデルのグラフィカル表現.

より現実の社会ネットワークに即したモデリングが可能となる．したがって，次数異質性を提案モデルのパラメータを用いて言い換えれば，たとえ両ノードが同じようなコミュニティ分布 (η_i, η_j) を持っていたとしても，ノードによってエッジの繋がりやすさが異なるということであり，まさしくエッジ確率パラメータである $\psi_{i\cdot\cdot}, \psi_{j\cdot\cdot}$ が次数異質性を表していると解釈できる．また，エッジ確率の $K \times K$ 行列表現は $\Psi_i = (\psi_{ikk'})$ で表され，行列の各要素は，事前分布としてベータ分布 $\psi_{ikk'} \mid \delta_{kk'}, \epsilon_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'})$ に従うことを仮定する．

本研究で用いるモデルは，上述の点以外は Igarashi and Terui (2020) の定式化を採用する．このとき，コミュニティ分布 H を所与としたときのネットワークデータに対する尤度は，式 (3.1) から次式に変更される．

$$
\begin{aligned}
(3.4)\ \ & p(A, S, R, \Psi \mid H) \\
& = p(A \mid S, R, \Psi) p(S \mid H) p(R \mid H) p(\Psi \mid \delta, \epsilon) \\
& = \prod_{i=1}^{D} \left\{ \prod_{j=1, j\neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi_j) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{ikk'} \mid \delta_{kk'}, \epsilon_{kk'}) \right\}.
\end{aligned}
$$

図 1 は提案モデルのグラフィカル表現である．

4. 条件付き事後分布とパラメータ推定

先行研究において，トピックモデルを推定するための手法は，変分ベイズ法や逐次学習法など多く提案されている．その中でも最も広く使われているものの一つが，崩壊型ギブスサンプリング (collapsed Gibbs sampling, CGS, Griffiths and Steyvers, 2004) である．これは，潜在変数の事後分布を導出する過程でモデルパラメータを積分消去し，サンプリングを効率的に行う手法である．以下では，本研究の提案モデルに対する CGS のための条件付き事後分布を導出する．

提案モデルにおける，コミュニティ分布 H，エッジ確率 Ψ，トピック分布 Θ，単語分布 Φ の

4 つのパラメータについては，事前分布との共役性に基づき，条件付き事後分布を既知の分布として導出することができる．ただし，その詳細な導出過程は付録 A に譲る．また，それ以外の潜在変数として，送り手及び受け手の潜在コミュニティ S, R，単語の潜在コミュニティ X 及び潜在トピック Z の 4 つがあるが，これらの条件付き事後分布は，付録 A で導出した事後分布を用いて以下のように導出される．

$$
\begin{aligned}
(4.1)\quad & p(s_{ij}=k, r_{ji}=k' \mid a_{ij}, A_{\setminus ij}, S_{\setminus ij}, R_{\setminus ji}, X, \gamma, \delta, \epsilon) \\
&\propto \int\int p(s_{ij}=k, r_{ji}=k' \mid \eta_i, \eta_j) p(x_i, x_j \mid \eta_i, \eta_j) p(\eta_i, \eta_j \mid S_{\setminus ij}, R_{\setminus ji}, X, \gamma) d\eta_i d\eta_j \\
&\quad \times \int p(a_{ij} \mid \psi_{jkk'}) p(\psi_{jkk'} \mid A_{\setminus ij}, S_{\setminus ij}, R_{\setminus ji}, \delta, \epsilon) d\psi_{jkk'} \\
&= \frac{N_{ik\setminus ij} + M_{ik} + \gamma_k}{\sum_t \left(N_{it\setminus ij} + M_{it} + \gamma_t\right)} \times \frac{N_{jk'\setminus ji} + M_{jk'} + \gamma_{k'}}{\sum_t \left(N_{jt\setminus ji} + M_{jt} + \gamma_t\right)} \times \\
&\quad \frac{\left(n^{(+)}_{jkk'\setminus ij} + \delta_{kk'}\right)^{\mathbb{I}(a_{ij}=1)} \left(n^{(-)}_{jkk'\setminus ij} + \epsilon_{kk'}\right)^{\mathbb{I}(a_{ij}=0)}}{n^{(+)}_{jkk'\setminus ij} + n^{(-)}_{jkk'\setminus ij} + \delta_{kk'} + \epsilon_{kk'}},
\end{aligned}
$$

$$
\begin{aligned}
(4.2)\quad & p(x_{im}=k, z_{im}=l \mid W, S, R, X_{\setminus im}, Z_{\setminus im}, \alpha, \beta, \gamma) \\
&\propto \int p(s_i, r_i \mid \eta_i) p(x_{im}=k \mid \eta_i) p(\eta_i \mid S, R, X_{\setminus im}, \gamma) d\eta_i \times \int p(z_{im}=l \mid \theta_k) \\
& p(\theta_k \mid X_{\setminus im}, Z_{\setminus,im}, \alpha) d\theta_k \times \int p(w_{im}=v \mid \phi_l) p(\phi_l \mid W_{\setminus im}, Z_{\setminus im}, \beta) d\phi_l \\
&= \frac{N_{ik} + M_{ik\setminus im} + \gamma_k}{\sum_t \left(N_{it} + M_{it\setminus im} + \gamma_t\right)} \times \frac{M_{kl\setminus im} + \alpha_l}{\sum_q \left(M_{kq\setminus im} + \alpha_q\right)} \times \frac{M_{lv\setminus im} + \beta_v}{\sum_u \left(M_{lu\setminus im} + \beta_u\right)}.
\end{aligned}
$$

ただし，式(4.1)における N_{ik} は，ノード i が持つ $D-1$ 個の関係性において，送り手及び受け手の潜在コミュニティとして k が割り当てられた回数を表し，M_{ik} は，ノード i の単語コミュニティに k が割り当てられた回数を表す．$n^{(+)}_{ikk'}$ は，ノード i に関する $D-1$ 個の関係性のうち，コミュニティ k, k' が割り当てられたエッジの数，$n^{(-)}_{ikk'}$ は，コミュニティ k, k' が割り当てられ，かつエッジのない関係性の数を表す．式(4.2)における M_{kl} は，コミュニティ k が割り当てられた単語のうちトピック l が割り当てられた回数，M_{lv} は，語彙 v にトピック l が割り当てられた回数を表す．また，添え字の $\setminus$ はこれらのカウントから，当該データを除くことを意味する．

CGS では，式(4.1)及び(4.2)に従って，各関係性及び単語に対して潜在コミュニティとトピックを繰り返しサンプリングする．最終的に，初期値に依存する稼働期間を除いたサンプルを用いて，積分消去していた 4 つのパラメータの期待値を計算することで推定値を得る．

5. 実証分析

5.1 使用データ

ここでは，現実のオンライン社会ネットワークに対して，提案モデルを用いた分析が有益であることを示すために，Twitter データを使った実証分析を行う．本節では，まず分析に用いたデータセットの概要と前処理について説明する．本研究では，任天堂株式会社が Twitter 上で保持している英語版公式アカウントを中心とするネットワークを対象として，以下の手順でデータを収集及び加工した．

表 2. WAIC によるモデル比較.

	$L=5$	$L=6$	$L=7$	$L=8$	$L=9$	$L=10$
$K=5$	4422206.32	4340879.93	4321068.95	4333535.35	4354814.11	4553144.83
$K=6$	4333313.32	4333488.66	4351008.38	4309479.01	4302773.27	4280703.13
$K=7$	4313265.58	4285253.01	**4272682.48**	4346780.91	4301005.75	4414800.13
$K=8$	4320416.87	4282485.37	4326300.05	4324393.23	4321806.29	4426226.19
$K=9$	4429170.84	4329997.66	4439594.82	4407656.85	4296128.61	4301655.85
$K=10$	4361219.83	4342899.53	4282056.30	4306509.44	4306244.12	4406655.34

まず，2018 年 5 月 1 日時点でのフォロー関係に従って，任天堂のアカウントをフォローしているユーザーからランダムにサンプリングを行った．続いて，サンプルされたユーザーをフォローしている別のユーザーからもランダムにサンプリングを行った．そして，それらのユーザーで形成されるネットワークにおいて，入次数と出次数の平均が 3 以下のユーザーを外れ値とみなしてデータセットから除外した．結果として，3,500 人のユーザーが残り，ネットワーク内におけるエッジの総数は 68,949 本であった．これらのユーザーで形成される有向グラフをネットワーク情報として使用する．

次に，テキストデータの作成方法を説明する．まず，上でサンプルされた 3,500 人分のアカウントに対して，2017 年 9 月 1 日から 2018 年の 2 月 28 日[1)] までに投稿した投稿内容からテキスト部分を全て抜き出した．これらのテキストデータに対して，文章から単語集合への分解，小文字への統一，数字，記号，及び主要なストップワード(a，the，I など)の削除，活用形から語幹への統一(stemming)の順に前処理を行った．さらに，処理済みのテキストデータのうち，コーパス内での頻度が 20 以下，あるいは 20 人以下のユーザーにしか使われていない低頻度の単語と，50 人以上のユーザーに使われている高頻度の単語を，トピック推定への悪影響を避けるためにデータセットから除いた．結果として，コーパス内には 9,001 種類の単語が残り，ノードごとの平均単語数は 98.2 であった．次節では，提案モデルにおけるコミュニティ数，トピック数の決定方法を説明したのち，作成したデータセットに対する提案モデルの推定結果について議論する．

5.2 分析結果

提案モデルを含めて，一般にブロックモデルを用いて分析する際には，事前にコミュニティ数(及び本研究ではそれに加えてトピック数)を決める必要がある．先行研究では，コミュニティ数の決定を情報量基準を用いたモデル比較として捉え，BIC による方法(Handcock et al., 2007; Saldaña et al., 2017)，integrated completed likelihood による方法(Daudin et al., 2008; Bouveyron et al., 2018)，変分ベイズによる方法(Latouche et al., 2012)など様々な手法が提案されている．しかし，本研究では，近年新たな情報量基準として提案され，現在では数多くの領域で使われている広く使える情報量基準(widely applicable information criterion, WAIC, Watanabe, 2010)をモデル比較の基準として採用した．提案モデルに対する WAIC の詳細は付録 B に譲る．表 2 は，コミュニティ数及びトピック数を 5 から 10 の範囲で設定し，5.1 節で作成したデータセットに対して WAIC を計算した結果である(K がコミュニティ数を，L がトピック数を表し，太字は表内で最少の値を意味する)．ただし，この時の繰り返し数は 5,000 回であり，そのうち 2,000 回を初期値に依存する稼働期間として除いた．また，ハイパーパラメータの設定は，それぞれ，$\alpha_l = 0.1, \forall l$，$\beta_v = 0.1, \forall v$，$\gamma_k = 1.0, \forall k$，$\delta_{kk'} = \epsilon_{kk'} = 0.1, \forall k, k'$ である．その結果，コミュニティ数 7，トピック数 7 のモデルが選ばれたため，以降ではこのモデルを用いた Twitter データの分析結果を議論する．

表 3. 提案モデルを用いて推定された単語分布において最も高い値を持つ上位 10 個の単語.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
nonfollow	teamemmmmsi	trapadr	criticalrol	iartg	growthhack	savvi
blackclov	dokkan	vevo	zeldathon	amread	digitalmarket	lube
hunterxhunt	twitchkitten	ddrive	orton	erotica	gdpr	foodporn
jojosbizarreadventur	vgc	leed	fursuitfriday	asmsg	smm	oiler
mkleosaga	roku	spinrilla	dramaalert	momlif	contentmarket	austria
wnf	wizebot	ifb	sdlive	hemp	gamedesign	tfc
hori	ryzen	gainwithpyewaw	htgawm	writerslif	podernfamili	crowdfir
mdva	freebiefriday	gainwithxtiandela	sml	bookreview	socialmedialmarket	tranc
hyrulesaga	streamersconnect	horford	robloxdev	kindleunlimit	bigdata	tock
nyxl	nbaliv	suav	yoongi	bookboost	emailmarket	texfil

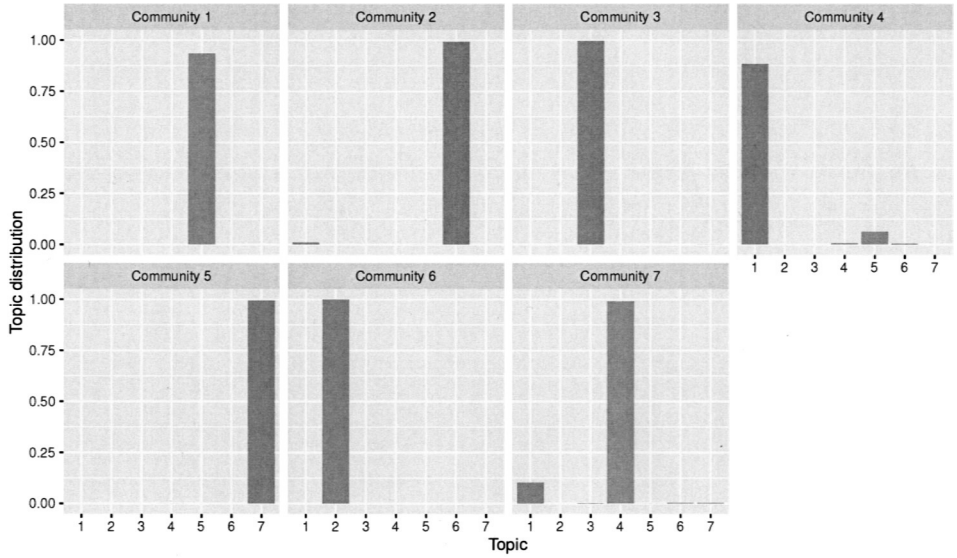

図 2. 提案モデルにおける各コミュニティのトピック分布に関する推定結果.

まず，ノードに依存しないグローバルパラメータ(単語分布 ϕ 及びトピック分布 θ)を見ることで，人々が検出されたコミュニティ内でどのようなことに関心を持っているのかが分かる．表 3 は，推定された単語分布の値が最も高い上位 10 個の単語をトピック毎に並べたものであり，これによってトピックの意味を解釈することができる．各トピックを代表する関連単語には下線が引かれており，トピックの意味は以下の通りに解釈できる[2)]．トピック 1 はアニメーションに関するトピック(代表的な単語は blackclov，hunterxhunt，jojosbizarreadventur など)，トピック 2 はストリーミング配信全般に関するトピック(代表的な単語は teamemmmmsi，twitchkitten，roku など)，トピック 3 は音楽に関するトピック(代表的な単語は vevo，spinrilla など)，トピック 4 はゲームストリーミング配信に関するトピック(代表的な単語は criticalrol，zeldathon など)，トピック 5 は読書に関するトピック(代表的な単語は amread，bookreview，kindleunlimit など)，トピック 6 はビジネスに関するトピック(代表的な単語は digitalmarket，smm，contentmarket など)，そしてトピック 7 はスポーツに関するトピック(代表的な単語は oiler，tfc など)と言える．

また，図 2 は，推定された各コミュニティのトピック分布であり，各コミュニティ内におけ

表 4. LDA を用いて推定された単語分布において最も高い値を持つ上位 10 個の単語.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
trapadr	teamemmmmsi	vevo	nonfollow	podernfamili	growthhack	gamedesign
ddrive	tfc	spinrilla	dokkan	iartg	digitalmarket	leed
ifb	twitchkitten	htgawm	zeldathon	amread	gdpr	savvi
gainwithpyewaw	hori	beck	criticalrol	asmsg	smm	lube
gainwithxtiandela	roku	orton	vgc	erotica	contentmarket	momlif
blackclov	mkleosaga	sdlive	fursuitfriday	foodporn	socialmediamarket	quoteoftheday
hunterxhunt	wnf	horford	dramaalert	dogsoftwitt	bigdata	austria
jojosbizarreadventur	wizebot	suav	sml	oiler	cto	hemp
yoongi	ryzen	herewego	robloxdev	writerslif	emailmarket	tranc
hoseok	streamersconnect	drippin	spforstreami	amiga	fintech	tock

るトピックの割合を確認することができる．これを見ると，コミュニティ毎のトピック分布が一つのトピックに集中しており，かつコミュニティ間で被りのない推定結果となっている．これは，提案モデルが，エッジが高密度かつテキストのトピックが類似するようなノード集合，つまりトピックベース・コミュニティを抽出するような構造になっているためと推察されるが，図 2 からのみでは判別できない．そこで，ネットワーク情報とテキスト情報を両方考慮する提案モデルの同時アプローチに対して，ネットワーク情報のみを考慮してコミュニティ構造を抽出する MMSB モデルと，テキスト情報のみを考慮してトピック構造を抽出する LDA モデルの結果を統合する独立アプローチとの比較を行うことで，図 2 の推定結果をさらに掘り下げていく．以下では，同時アプローチと独立アプローチに対して，単語分布によるトピックの解釈，コミュニティごとのトピック分布，推定されるコミュニティ構造をそれぞれ比較する．

まず，表 4 に，LDA によって抽出されたトピックに関する代表的な単語が示されている．これを見ると，提案モデルの推定結果である表 3 と同じ単語が同一のトピックに多く並んでおり，ネットワークとテキストを両方考慮したモデリングと，テキストのみのモデリングで同一のトピックを抽出していることが確認できる．

次に，MMSB モデルと LDA の結果を統合してコミュニティ毎のトピック分布を評価する．LDA モデルは，文書を単語集合とみなして文書ごとのトピック分布を推定するモデルであるが，ここではノードに単語集合が付随するとみなすため，ノードごとにトピック分布が推定される．また，MMSB モデルによってもノードごとにコミュニティへの所属割合が推定されている．したがって，コミュニティ所属割合で重みづけてトピック分布を足し合わせることで，提案モデルで推定されるようなコミュニティごとのトピック分布を事後的に導出することができる．LDA モデルによって推定されたノードごとのトピック分布を $\hat{\lambda}_i^{(ind)}$，MMSB モデルによって推定されたノードごとのコミュニティ分布を $\hat{\eta}_i^{(ind)}$ とすると，独立アプローチにおけるコミュニティごとのトピック分布 $\theta_k^{(ind)}$ は以下で導出される．

$$\theta_k^{(ind)} = \sum_{i=1}^{D} \hat{\lambda}_i^{(ind)} \times \hat{\eta}_{ik}^{(ind)}, \qquad k = 1, \ldots, K. \tag{5.1}$$

その結果は図 3 に示され，提案モデルの推定結果とは大きく異なり，一つのコミュニティに複数のトピックが対応していることが分かる．

そして，次数異質性を考慮したネットワークのみのモデルと提案モデルのコミュニティ内エッジ密度を比較する．両モデルとも相手ノードによって所属コミュニティが異なる混合メンバーシップを仮定しているため，推定されたコミュニティ分布に従って，ノードごとに最も高い値を持つコミュニティに所属するとしてエッジ密度を計算する．両モデルのエッジ密度及び

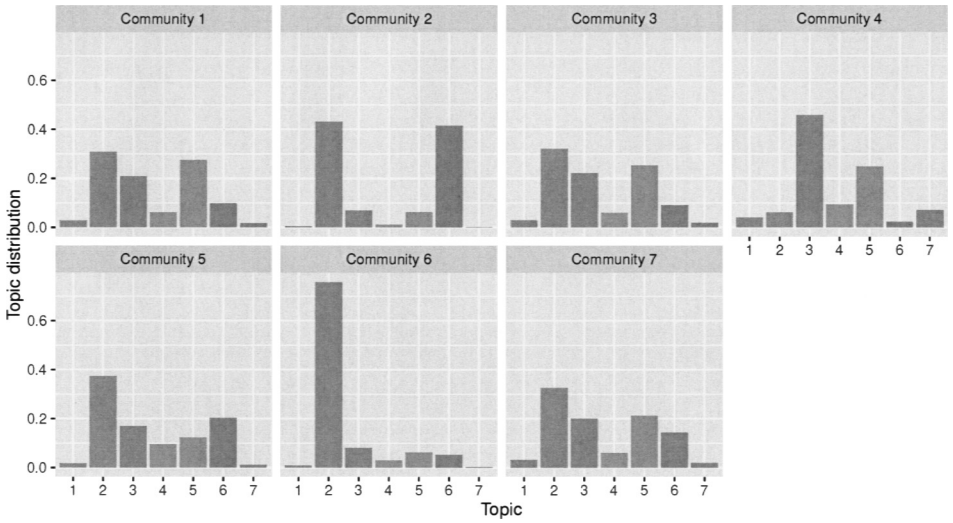

図 3. ネットワーク情報のみを考慮した比較モデルにおける各コミュニティのトピック分布.

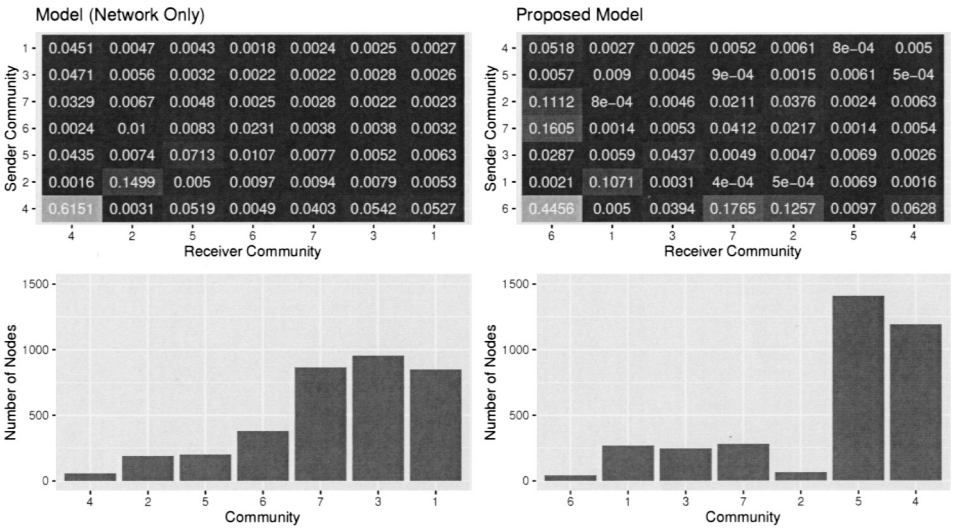

図 4. 提案モデル及び比較モデルにおける各コミュニティ内のエッジ密度と各コミュニティに所属するノード数.

コミュニティを構成するノードの数が図 4(左上と左下の図がネットワークのみを扱う MMSB モデルによる結果であり，右上と右下の図が提案モデルによる結果である)に示されている．なお，比較のため，コミュニティ内エッジ密度(対角成分)の大きい順にコミュニティ番号を並び変えている．これを見ると，多くのノードが所属するエッジ密度の低いコミュニティについて，ネットワークのみのモデルでは三つ，提案モデルでは二つで推定しているため，両モデルで値が全体的に少し異なっている．しかし，それ以外のコミュニティ構造については両者とも同様の構造を捉えており，例えば，少数のノードで構成されているコミュニティ(ネットワークのみのモデルでは 4 番，提案モデルでは 6 番のコミュニティ)や，他コミュニティとの関係

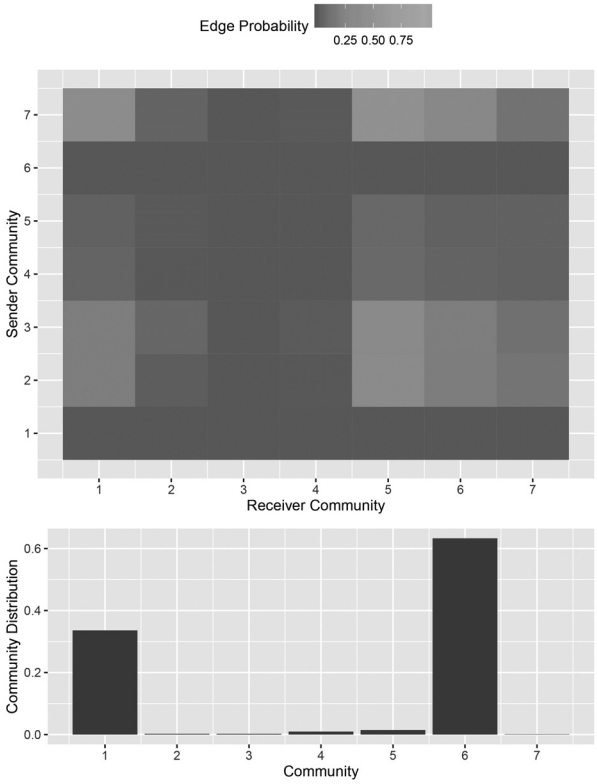

図 5. ノード 1 のエッジ確率とコミュニティ分布に関する推定結果.

は希薄であるものの，内部では比較的高い密度でつながっている中規模のコミュニティ(ネットワークのみモデルにおける 2，5 番，提案モデルにおける 1，3 番のコミュニティ)などが共通して推定されている．

以上により，提案モデルは，ネットワークあるいはテキストのみを考慮するモデルと全体的には同様のコミュニティ構造及びトピック構造を捕捉しながら，コミュニティ内のトピック構造をより明確に表現するモデルであると言える．ただし，これは今回用いたデータセットにおける結果であり，一般のネットワークにおける性質を検証するには理論解析など，さらなる議論が必要である．

最後に，各ノードについて異質なローカルパラメータ(エッジ確率の ψ)の推定結果を確認する．図 5 及び図 6 は，ノード番号 1 番と 237 番に関するエッジ確率とコミュニティ分布の推定結果である．また，ノード 1 の入次数は 6，出次数は 0 であり，ノード 237 の入次数は 657，出次数は 37 である．推定結果は，この両ノードの次数異質性を如実に表しており，ノード 1 が主に属するコミュニティ(コミュニティ 1 と 6)に関するエッジ確率は低い値で推定されているのに対して，ノード 237 が主に属するコミュニティ(コミュニティ 1 と 5)に関するエッジ確率は高い値で推定されている．このように，エッジ確率のパラメータに次数異質性を考慮した仮定を導入することで，より柔軟にネットワークモデルを表現できるようになり，テストデータに対する予測性能も向上することが期待される．次節では，これを検証するために，提案モデルからテキストの活用及び次数異質性の考慮という特徴を除いた比較モデルと共に比較実験を

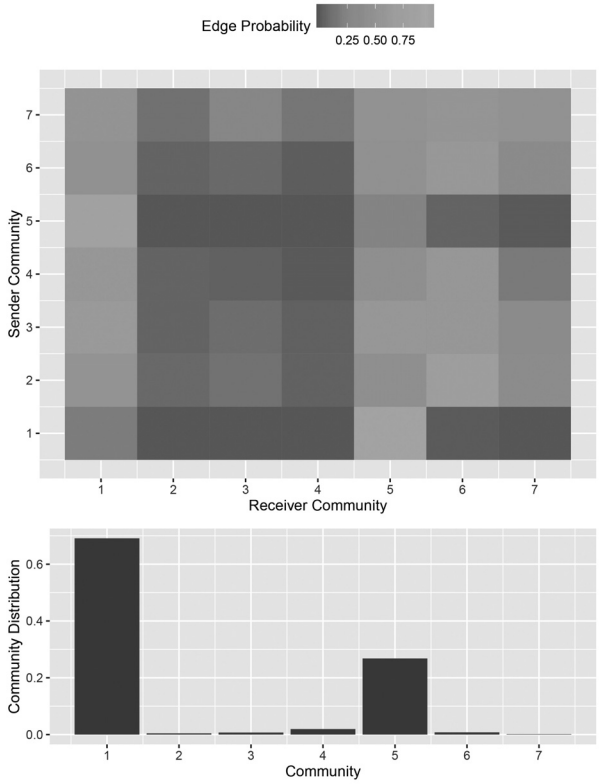

図 6. ノード 237 のエッジ確率とコミュニティ分布に関する推定結果.

行う.

5.3 予測性能の検証

本節では，提案モデルのテストデータに対する予測性能及びテキスト情報の活用と次数異質性の考慮が予測性能に与える効果を比較モデルとの比較を通して検証する．提案モデル（以降モデル IV）がテキスト情報の活用と次数異質性の考慮をどちらも含んだモデルであるのに対して，比較モデルとして，テキスト情報を活用せずかつ次数異質性を考慮しないモデル（以降モデル I），テキスト情報を活用せず次数異質性を考慮するモデル（以降モデル II），そしてテキスト情報を活用するが次数異質性は考慮しないモデル（以降モデル III）の 3 種類を検討する．

5.2 節では，全てのネットワーク，テキストデータを学習データとしてモデルの推定を行ったが，ここでは，各ノードが持つ $D-1$ 個の関係性のうち，90% を学習データとしてモデルの推定に使い，残りの 10% をテストデータとした．テキストデータについては，前節同様全てのデータを学習データとして用いた．また，繰り返し数やハイパーパラメータの設定も前節と同じ条件で推定している．これらの条件の下で学習データに対する推定を行い，各パラメータの推定値を得た．推定されたコミュニティ分布とエッジ確率を $\hat{H}, \hat{\Psi}$ と表すと，例えば提案モデルについては，テストデータ $a_{ij} \in A^{test}$ に対する予測確率は以下で計算できる．

$$p(a_{ij}=1) = \sum_{k=1}^{K}\sum_{k'=1}^{K} \hat{\eta}_{ik}\hat{\eta}_{jk'}\hat{\psi}_{jkk'} \tag{5.2}$$

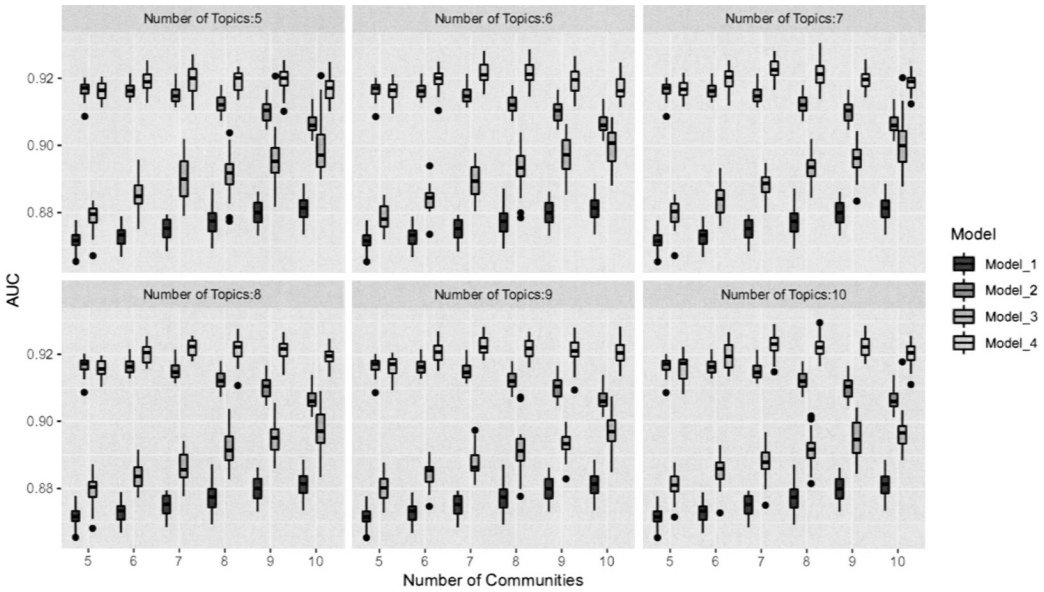

図 7. AUC による予測性能の比較.

3 種類の比較モデルについても同様に，コミュニティ分布とエッジ確率の積によって予測確率を計算できる．

ここでは，コミュニティ数とトピック数をそれぞれ 5 から 10 の値で設定し，各モデルの Area Under the Curve(AUC)をグリッド状に計算した．また，各モデル，各コミュニティ数及びトピック数について，テストデータを入れ替えながらモデルの推定と AUC の計算を 30 回繰り返した．図 7 はその時の AUC をまとめた箱ひげ図である．

まず，次数異質性を考慮しないモデル(モデル I，モデル III)と考慮するモデル(モデル II，IV)の比較を行う．次数異質性を考慮しない場合，コミュニティ数の増加と共に AUC が上昇している．一般的に，最適なクラスタ数に達するまでは，クラスタ数を増やすほどクラスタリングの精度は上昇するため，これは期待通りの結果と言える．一方で，次数異質性を考慮するモデルは，全体的に AUC が高く，異質性を導入することで，ネットワークモデルとしての予測性能が上がることが示されている．しかし，モデル II ではコミュニティ数の増加と共に AUC が下降しており，次数異質性が予測性能を悪化させる側面も持ち合わせている可能性がある．これは，異質性を導入することでノード一つ一つのエッジ確率を推定するためのデータが少なくなるため，コミュニティ数の増加と共にその影響が顕在化していると推察され，本研究で提案する形で異質性を導入するデメリットとも言える．その場合，階層ベイズを用いるなどして，エッジ確率の異質性を作り出す要因としてノード間に共通する構造を取り込むことで改善の可能性がある．

次に，テキスト情報を考慮しないモデル(モデル I，モデル II)と考慮するモデル(モデル III，モデル IV)を比較すると，テキスト情報を考慮することで予測精度が向上していることが分かる．さらに，その効果はコミュニティ数に対して一定ではなく，コミュニティ数が増えるにしたがって，テキスト効果による予測精度の上昇幅は大きくなっている．つまり，ネットワーク上のコミュニティを細かく分割していく際に，ネットワーク情報だけでは分割が難しい場面であっても，テキスト情報を使うことで，外挿予測に対しても頑健な，より精緻なクラスタリン

グが可能となる．提案モデルであるモデル IV に関しては，前述した次数異質性による予測性能への悪影響が予想されるが，テキスト効果による予測性能の改善が相殺しあうことで，コミュニティ数が増えても予測性能が落ちることなく，高い AUC を維持していると考えられる．また，トピック数については，その数が変化しても各モデル（モデル III，IV）の予測性能には大きな影響を与えていない．

結果として，テキスト情報と次数異質性を考慮する提案モデルがそれぞれの効果を考慮しない比較モデルとの比較の中で，最も予測性能に優れたモデルであることが示された．ただし，解決されていない課題もあり，次数異質性のモデルへの反映させ方によっては予測性能を悪化させる可能性があることについては，モデルの改善も含めて今後の課題としたい．

6. おわりに

本研究では，社会ネットワーク分析をより現実に即した有意義な分析とするために，ネットワーク情報だけでなく，人々の興味や関心を表すソーシャルメディア上のテキスト情報を考慮し，さらに，社会ネットワークに特有の次数異質性を加味したモデルを提案した．先行研究における既存モデルと比較したとき，本研究で提案するモデルの特色としては，ネットワーク上の各ノードが持つテキスト情報を利用している点，ノードがそれぞれの関係性によって異なるコミュニティに属することを許容している点，無向グラフか有向グラフにかかわらず適用できる点，そして次数異質性を考慮し，エッジ確率のパラメータがノードごとに異質であることを仮定している点が挙げられる．これによって，次数がノードによって大きく異なる一般的な社会ネットワークに対しても十分な適合性能を有しながら，エッジが密に集まっており，かつノードごとに固有のテキストデータが同一の分布から生成される，トピックベース・コミュニティの検出が可能となる．

実証分析の結果，崩壊型ギブスサンプリングによって推定される提案モデルは，現実の Twitter データに対して，解釈可能なコミュニティ及びトピック構造を捉えるだけでなく，次数異質性やテキスト情報を考慮しない比較モデルよりも優れた予測性能を持っていることが示された．さらに，この結果から，オンラインの社会ネットワーク分析において，ネットワーク上の大まかなコミュニティ構造を超えて，さらに細かくクラスターを分析していく場合は，各ノードが持つテキスト情報を加味することが外挿予測を向上させるという点で有益であることが分かった．

本研究では，オンライン上の社会ネットワークに着目したため，人々がネットワークを形成する際には，相手のネットワーク情報とテキスト情報のみを考慮するという仮定を置き，相手の年齢や性別といった属性情報，あるいは行動や態度といった情報は，これらのデータが利用できないことから提案モデルの考慮から外していた．一方で，社会ネットワーク分析に関する先行研究の文脈では，そのようなノード固有の（あるいはノード間の）特徴量がネットワーク形成に影響していることが多くの研究で示されている（Hoff et al., 2002; Handcock et al., 2007）．本研究では，エッジ形成の関数が，関係性を結ぶ両者のコミュニティ分布，及び関係性を受け取る側のエッジ確率で構成されていたが，先行研究を参照すると，ここに属性や行動情報といったノード固有の特徴量を組み込む拡張は有意義であり，これらの情報をモデルに取り込むことは直接的に可能である．データの利用可能性と合わせて今後の課題としたい．

注．

[1] テキストデータの前処理の段階で，大半のユーザーが，2018 年 3 月に開かれた Nintendo Direct という新商品発表イベントに関する投稿を行っていることが判明した．したがっ

て，本研究では，このような多くのユーザーで共通する同一の事象に対する投稿がトピックの推定に与える影響を避けるため，テキストデータの観測期間を 2018 年 2 月 28 日までとした．

[2)] 本文中で挙げた各トピックの代表的な単語の解説は次の通りである．トピック 1 の単語は，それぞれ，ブラッククローバー，ハンターハンター，ジョジョの奇妙な冒険という著名なアニメ・漫画作品のタイトルである．トピック 2 の単語は，配信者のグループ名や配信プラットフォーム名である．トピック 3 の単語は，ミュージックビデオなどのコンテンツを配信しているサイト名である．トピック 4 の単語は，ダンジョンズ&ドラゴンズやゼルダの伝説など特定のゲームをプレイし配信するプロジェクトである．トピック 5 の単語は，読書に関する Twitter 上のハッシュタグである．トピック 6 の単語は，オンラインマーケティングに関する情報を発信する際のハッシュタグである．トピック 7 の単語は，データセットの観測期間中に話題になったスポーツ大会の名前やアイスホッケーのチーム名である．

謝　辞

本研究は JSPS 科研費 18J20698 及び(A)17H01001 の助成を受けている．改訂に際して，査読者より大変有意義なコメントを頂戴した．

付　録

A. 条件付き事後分布の導出

3 節では，式(4.1)及び(4.2)で潜在コミュニティ及び潜在トピックの条件付き事後分布を導出した．これらの事後分布を得るためには，まず，コミュニティ分布，エッジ確率，トピック分布，単語分布の 4 つのパラメータについて，条件付き事後分布を導出する必要がある．事前分布との共役性に基づいて，これらの事後分布は以下のように導出される．

$$p(\eta_i \mid S, R, X, \gamma) = \frac{\Gamma\left(\sum_k N_{ik} + M_{ik} + \gamma_k\right)}{\prod_k \Gamma(N_{ik} + M_{ik} + \gamma_k)} \prod_{k=1}^{K} \eta_{ik}^{N_{ik}+M_{ik}+\gamma_k} \tag{A.1}$$

$$p(\psi_{ikk'} \mid A, S, R, \delta, \epsilon) = \frac{\Gamma(n_{ikk'}^{(+)} + n_{ikk'}^{(-)} + \delta_{kk'} + \epsilon_{kk'})}{\Gamma(n_{ikk'}^{(+)} + \delta_{kk'})\Gamma(n_{ikk'}^{(-)} + \epsilon_{kk'})} \times \psi_{ikk'}^{\mathbb{I}(a_{ij}=1)}(1 - \psi_{ikk'})^{\mathbb{I}(a_{ij}=0)} \tag{A.2}$$

$$p(\theta_k \mid X, Z, \alpha) = \frac{\Gamma\left(\sum_l M_{kl} + \alpha_l\right)}{\Pi_l \Gamma(M_{kl} + \alpha_l)} \prod_{l=1}^{L} \theta_{kl}^{M_{kl}+\alpha_l} \tag{A.3}$$

$$p(\phi_l \mid W, Z, \beta) = \frac{\Gamma\left(\sum_v M_{lv} + \beta_v\right)}{\Pi_v \Gamma(M_{lv} + \beta_v)} \prod_{v=1}^{V} \phi_{lv}^{M_{lv}+\beta_v}. \tag{A.4}$$

B. 提案モデルに対する WAIC の定義式

提案モデルに対する WAIC の定義式は以下の通りである．

$$WAIC = -2\sum_{i=1}^{D}(lpd^{(i)} - p_{waic}^{(i)}), \tag{B.1}$$

$$lpd^{(i)} = \log\left(\frac{1}{G}\sum_{g=b+1}^{G}\prod_{j=1}^{D} p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}) \prod_{m=1}^{M_i} p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)})\right)$$

$$p_{waic}^{(i)} = \frac{G}{G-1}\left(\frac{1}{G}\sum_{g=b+1}^{G}\left(\sum_{j=1}^{D}\log p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)})^2 + \sum_{m=1}^{M_i}\log p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)})^2\right)\right.$$
$$\left. - \left(\frac{1}{G}\sum_{g=b+1}^{G}\left(\sum_{j=1}^{D}\log p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}) + \sum_{m=1}^{M_i}\log p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)})\right)\right)^2\right).$$

ただし，$p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)})$ と $p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)})$ は，CGS によるサンプルのうち g 回目の繰り返しにおけるサンプルで推定したパラメータを用いて計算される尤度であり，以下で定義されている．

$$p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}) = \sum_{k=1}^{K}\sum_{k'=1}^{K} \eta_{ik} \cdot \eta_{jk'}^{(g)} \cdot \psi_{jkk'}^{(g)\mathbb{I}(a_{ij}=1)} \cdot (1-\psi_{jkk'})^{(g)\mathbb{I}(a_{ij}=0)} \tag{B.2}$$

$$p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)}) = \sum_{k=1}^{K}\sum_{l=1}^{L} \eta_{ik}^{(g)} \cdot \theta_{kl}^{(g)} \cdot \phi_{lw_{im}}^{(g)}. \tag{B.3}$$

参 考 文 献

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.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research*, **3**(4-5), 993–1022.

Bouveyron, C., Latouche, P. and Zreik, R. (2018). The stochastic topic block model for the clustering of vertices in networks with textual edges, *Statistics and Computing*, **28**(1), 11–31.

Chang, J. and Blei, D. M. (2010). Hierarchical relational models for document networks, *The Annals of Applied Statistics*, **4**(1), 124–150.

Daudin, J.-J., Picard, F. and Robin, S. (2008). A mixture model for random graphs, *Statistics and Computing*, **18**(2), 173–183.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics, *Proceedings of the National Academy of Sciences*, **101**(Supplement 1), 5228–5235.

Handcock, M. S., Raftery, A. E. and Tantrum, J. M. (2007). Model-based clustering for social networks, *Journal of the Royal Statistical Society. Series A*: *Statistics in Society*, **170**(2), 301–354.

Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis, *Journal of the American Statistical Association*, **97**(460), 1090–1098.

Igarashi, M. and Terui, N. (2020). Characterization of topic-based online communities by combining network data and user generated content, *Statistics and Computing*, DOI: http://dx.doi.org/10.1007/s11222-020-09947-5.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks, *Physical Review E — Statistical, Nonlinear, and Soft Matter Physics*, **83**(1), 016107.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E. and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models, *Social Networks*, **31**(3), 204–213.

Latouche, P., Birmelé, E. and Ambroise, C. (2012). Variational Bayesian inference and complexity control for stochastic block models, *Statistical Modelling*: *An International Journal*, **12**(1), 93–115.

Liu, Y., Niculescu-Mizil, A. and Gryc, W. (2009). Topic-link LDA: Joint models of topic and author community, *Proceedings of the 26th International Conference on Machine Learning, ICML 2009*, 665–672.

Saldaña, D. F., Yu, Y. and Feng, Y. (2017). How many communities are there?, *Journal of Computational and Graphical Statistics*, **26**(1), 171–181.

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure, *Journal of Classification*, **14**(1), 75–100.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs, *Journal of the American Statistical Association*, **82**(397), 8–19.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research*, **11**, 3571–3594.

Zhu, Y., Yan, X., Getoor, L. and Moore, C. (2013). Scalable text and link analysis with mixed-topic link models, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 473–481.

A Model for Social Network Analysis Considering Text Information on Social Media —Extended Model Considering Node Degree Heterogeneity—

Mirai Igarashi and Nobuhiko Terui

Graduate School of Economics and Management, Tohoku University

In the study of social networks, it has become increasingly important to consider not only network information but also text information generated by people on social media in order to deeply understand the community structure. By taking text information into account, it is possible to analyze social networks with complex community structures, which have multiple clusters of people depending on their interests in a single network with densely connected edges. In this study, we extend the existing model that simultaneously consider network and text information, and propose a model with degree heterogeneity to represent that the probability of edge generation varies for each node. In empirical analysis using Twitter dataset, we compare it with comparative models not using text information or degree heterogeneity, and show it's better predictive performance.

Key words: Social network analysis, community detection, text analysis, topic modeling, Bayes inference, node heterogeneity.