DSSR

Discussion Paper No.J-7

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

五十嵐未来 照井伸彦

2020 年 5 月

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

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

五十嵐未来 *　　照井伸彦 †

2020 年 5 月

Abstract

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 Igarashi and Terui (2020) 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，エッジ生成確率を関係するノードごとに異なるように定式化することで，ノードの次数分布がべき乗則に従うという社会ネットワークが持つ一般的な性質を考慮するモデルを提案している．Twitter を用いた実証分析では，提案モデルを用いた実データへの応用例を示すとともに，先行研究における既存モデルよりも優れた予測性能を持つことを示す．

Keywords: 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・ノード次数の異質性

*東北大学大学院経済学研究科　博士後期課程：〒 980-8576　宮城県仙台市青葉区川内 27 − 1（E-mail : mirai.igarashi.s7@dc.tohoku.ac.jp）．本研究は JSPS 科研費 18J20698 の助成を受けたものです．

†東北大学大学院経済学研究科　教授（E-mail : terui@tohoku.ac.jp）．本研究は JSPS 科研費 (A) 17H01001 の助成を受けたものです．

1　序論

Social Networking Sites (SNS) の流行や e-コマース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，ネットワーク構造を捉え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，そこからコミュニティ構造を抽出することを目的としている．また，社会ネットワークにおいて，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 e-コマースサイトの台頭などにより，ユーザー生成コンテンツ (User-Generated-Contents，UGC)，特にテキスト情報をネットワーク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としては，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抽出できるという点が挙げられる．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．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，ネットワークとテキストを考慮したモデルによる検出を提案している．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，現実世界における社会的なつながりだけでなく，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ため，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することで，社会ネットワーク分析モデルを精緻化させることができる．

また，社会ネットワークが持つ性質の一つとして，次数分布がべき乗則に従うという性質がある．これは，ごく少数の人々が多くの人々とネットワーク上で関係を持ち，その他大勢の人々は，ごく少数の人々とのみ関係性を持つ傾向にあるという性質である．そのように，現実の社会ネットワークにおいて，次数はノードごとに異質であるが，確率的ブロックモデル (Snijders and Nowicki, 1997) など代表的なネットワークモデルの多くがそうであるように，Igarashi and Terui (2020) では，次数の異質性を考慮していない．

本研究では，Igarashi and Terui (2020) のモデルを拡張し，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質なエッジ確率とし，次数の異質性を考慮したモデルを提案する．本研究の実証分析では，次数の異質性を考慮した提案モデルを，一般的な確率ブロックモデルと同様に異質性を考慮しない差分モデルと比較し，外装予測において提案モデルの方が優れていることを示す．

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，モデルが提案されて以降，様々な文脈でモデルの拡張が取り組まれており，例えば，Airoldi et al. (2008) は，SBM の拡張モデルとして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．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．MMSB では，各ノードは，他ノードとの関係性毎に複数のコミュニティに属することが許容されている．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}，ノード j が属するコミュニティを r_{ji} 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．これらの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ことであるのに対して，本研究の目的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ことであるように対照的なものである．

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ている部分は本研究と異なる点である．また，Liu et al. (2009) では，エッジ生成確率が，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，対象とするネットワークを無向グラフとして扱うこと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ことで拡張した，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．この点において本研究における提案モデルとも似た構造を有しているが，主な相違点は，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る．言い換えれば，Zhu et al. (2013) は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といえる．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メンバーが密なリンク構造を持ってい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，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，Igarashi and Terui (2020) では，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提案している．本研究では，このモデルを拡張し，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分布の異質性を考慮したモデリングが可能となる．先行研究においては，Karrer and Newman (2011) が，SBM で定義されるようなノードについて同質的なエッジ生成確率を適用するのではなく，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．本研究における定式化では，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定義しており，Karrer and Newman (2011) とも異なるアプローチをとっている．

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の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色と言える．これらの比較を通して，5 節では，提案モデルから次数の異質性の考慮を除いたモデルにあたる Igarashi and Terui (2020)，及びテキスト情報を考慮しないモデルに相当する Airoldi et al. (2008) を比較モデルとしてそれらの予測性能を検証している．

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) のモデルを紹介し，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 は潜在コミュニティ $r_{ji} \in \{1, \ldots, K\}$ 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 の事前分布はディリクレ分布 $\eta_i \mid \gamma \sim Dirichlet(\gamma)$ 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，$a_{ij} \mid s_{ij} = k, r_{ji} = k', \Psi \sim Bernoulli\,(\psi_{kk'})$ 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，$\psi_{kk'} \mid \delta_{kk'}, \epsilon_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'})$ を持つ．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．

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

続いて，ノード固有のテキストコンテンツについて考える．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．

$$
\begin{aligned}
&p(W, X, Z, \Theta, \Phi \mid H) \\
&= p(W \mid Z, \Phi) p(Z \mid X, \Theta) p(X \mid H) p(\Theta \mid \alpha) p(\Phi \mid \beta) \\
&= \prod_{i=1}^{D} \left\{ \prod_{m=1}^{M_i} \{ p(w_{im} \mid z_{im}, \Phi) p(z_{im} \mid x_{im}, \Theta) p(x_{im} \mid \eta_i) \} \right\} \prod_{k=1}^{K} p(\theta_k \mid \alpha) \prod_{l=1}^{L} p(\phi_l \mid \beta). \qquad (2)
\end{aligned}
$$

コミュニティ分布 H を所与とすることで，式（1）及び（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（1）と（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．

$$
\begin{aligned}
&p(A, W, S, R, X, Z, H, \Psi, \Theta, \Phi) \\
&= \prod_{i=1}^{D} \left\{ \prod_{j=1, j \neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \prod_{m=1}^{M_i} \{ p(w_{im} \mid z_{im}, \Phi) P(z_{im} \mid x_{im}, \Theta) p(x_{im} \mid \eta_i) \} \right\} \times \\
&\prod_{i=1}^{D} p(\eta_i \mid \gamma) \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{kk'} \mid \delta_{kk'}, \epsilon_{kk'}) \prod_{k=1}^{K} p(\theta_k \mid \alpha) \prod_{l=1}^{L} p(\phi_l \mid \beta). \qquad (3)
\end{aligned}
$$

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フィッティング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 $a_{ij} = 1 \mid s_{ij} = k, r_{ji} = k' \sim Bernoulli(\psi_{jkk'})$ 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，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．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，上述した点以外は Igarashi and Terui (2020) と同じ定式化を仮定しているため，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，式 (1) が以下のように変更される．

$$
\begin{aligned}
& p(A, S, R, \Psi \mid H) \\
& = p(A \mid S, R, \Psi) p(S \mid H) p(R \mid H) p(\Psi \mid \delta, \epsilon) \\
& = \prod_{i=1}^{D} \left\{ \prod_{j=1, j \neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi_j) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{ikk'} \mid \delta_{kk'}, \epsilon_{kk'}) \right\}. \qquad (4)
\end{aligned}
$$

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 Appendix 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，Appendix

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．

$$
\begin{aligned}
&p(s_{ij}=k, r_{ji}=k' \mid a_{ij}, A_{\backslash ij}, S_{\backslash ij}, R_{\backslash ji}, X, \gamma, \delta, \epsilon) \\
&\propto \int\int p(s_{ij}=k \mid \eta_i) p(r_{ji}=k' \mid \eta_j) p(x_i \mid \eta_i) p(x_j \mid \eta_j) p(\eta_i \mid S_{\backslash ij}, R_{\backslash ji}, X, \gamma) \\
&\quad p(\eta_j \mid S_{\backslash ij}, R_{\backslash ji}, X, \gamma) d\eta_i d\eta_j \times \int p(a_{ij} \mid \psi_{jkk'}) p(\psi_{jkk'} \mid A_{\backslash ij}, S_{\backslash ij}, R_{\backslash ji}, \delta, \epsilon) d\psi_{jkk'} \\
&= \frac{N_{ik\backslash ij} + M_{ik} + \gamma_k}{\sum_t \left(N_{it\backslash ij} + M_{it} + \gamma_t\right)} \times \frac{N_{jk'\backslash ji} + M_{jk'} + \gamma_{k'}}{\sum_t \left(N_{jt\backslash ji} + M_{jt} + \gamma_t\right)} \times \\
&\quad \frac{\left(n^{(+)}_{jkk'\backslash ij} + \delta_{kk'}\right)^{\mathbb{I}(a_{ij}=1)} \left(n^{(-)}_{jkk'\backslash ij} + \epsilon_{kk'}\right)^{\mathbb{I}(a_{ij}=0)}}{n^{(+)}_{jkk'\backslash ij} + n^{(-)}_{jkk'\backslash ij} + \delta_{kk'} + \epsilon_{kk'}},
\end{aligned}
\tag{5}
$$

$$
\begin{aligned}
&p(x_{im}=k, z_{im}=l \mid W, S, R, X_{\backslash im}, Z_{\backslash im}, \alpha, \beta, \gamma) \\
&\propto \int p(s_i, r_i \mid \eta_i) p(x_{im}=k \mid \eta_i) p(\eta_i \mid S, R, X_{\backslash im}, \gamma) d\eta_i \times \int p(z_{im}=l \mid \theta_k) \\
&\quad p(\theta_k \mid X_{\backslash im}, Z_{\backslash, im}, \alpha) d\theta_k \times \int p(w_{im}=v \mid \phi_l) p(\phi_l \mid W_{\backslash im}, Z_{\backslash im}, \beta) d\phi_l \\
&= \frac{N_{ik} + M_{ik\backslash im} + \gamma_k}{\sum_t \left(N_{it} + M_{it\backslash im} + \gamma_t\right)} \times \frac{M_{kl\backslash im} + \alpha_l}{\sum_q \left(M_{kq\backslash im} + \alpha_q\right)} \times \frac{M_{lv\backslash im} + \beta_v}{\sum_u \left(M_{lu\backslash im} + \beta_u\right)}.
\end{aligned}
\tag{6}
$$

ただし，式（5）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が生成されている関係性の数，$n^{(-)}_{ikk'}$ は，エッジが生成されていない関係性の数を表す．式（6）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 $\backslash$ 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（5）及び（6）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．

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

[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．

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 applecable 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 Appendix 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．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．

まず，ノードに依存しないグローバルパラメータを見ることで，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．図 2 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．各トピックの意味と代表的な単語は以下の通りである．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，sprinrilla 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 critical role，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 3 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るトピックの割合を確認することができる．

次に，各ノードについて異質なローカルパラメータの推定結果を確認する．図 4 及び 5 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中心性の違い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がエッジの繋がりやすさに関する異質性を捉えられるような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，先行研究における既存モデルと共に比較実験を行う．

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を，比較モデルと共に検証する．比較モデルとして，先行研究における既存モデルから，Airoldi et al. (2008) と Igarashi and Terui (2020) を選んだ．Airoldi et al. (2008) のモデルは，Igarashi and Terui (2020) のモデルからテキスト情報の考慮を除いたモデルに相当するため，これらを比較することで，テキスト情報を考慮することによる予測性能への影響を見ることができる．さらに，Igarashi and Terui (2020) のモデルは，本研究のモデルからエッジ確率の異質性を除いた同質モデルであり，異質性の構造が予測性能へ与える影響を見ることができる．

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しした．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．

$$p(a_{ij} = 1) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \hat{\eta}_{ik} \hat{\eta}_{jk'} \hat{\psi}_{jkk'} \tag{7}$$

Airoldi et al. (2008) と Igarashi and Terui (2020) の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．

表 3 は，コミュニティ数とトピック数をそれぞれ 5 から 10 まで変化させたときの各モデルの Area Under the Curve（AUC）の値である．これを見ると，ほぼ全ての組み合わせについて提案モデル（表内では Hetero）が比較モデルである Igarashi and Terui (2020)（表内では Homo）よりも優れていることが分かる．よって，社会ネットワークに一般的にみられる次数の異質性を考慮し，エッジが生成される確率はノードごとに同質的ではないという仮定を置いたモデルの方が予測性能が優れたモデルであるといえる．また，Airoldi et al. (2008)（表内では MMSB）と Igarashi and Terui (2020) を比較すると，コミュニティの数が少ないとき（$K = 5, 6$）には Airoldi et al. (2008) の方が AUC が全体的に高く，コミュニティの数が多いとき（$K = 8, 9, 10$）は Igarashi and Terui (2020) の方が全体的に AUC が高いという結果であった．この結果から，本研究で用いた Twitter ネットワークは，大まかにコミュニティを分ける際にはネットワーク情報のみを考慮するだけで充分であるが，より細かなコミュニティに分ける場合には，テキスト情報を用いてトピックのまとまりを考慮しながら分けた方がより良いクラスタリングとなるネットワークであるといえる．つまり，1 節でも述べたように，学校や同級生といった大きなまとまりのコミュニティの中から，音楽やスポーツのように趣味や関心事が共通しているまとまり，トピックベース・コミュニティを見つけていくモデリングが，より精緻なネットワーク分析のためには有益であるという示唆が得られた．

6　結論

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の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沿って複数の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，そして次数の異質性を考慮し，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フィッティング性能を有しながら，エッジが密に集まっており，かつ生成されたテキストのトピック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，意味のあるコミュニティ及びトピック構造を捉えるだけでなく，どのようなコミュニティ数，トピック数の組み合わせであっても既存モデルよりも優れた予測性能を持っており，次数の異質性を考慮してエッジ確率を推定するモデリングは，予測性能において優れ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してトピックベース・コミュニティを見つけていくモデリングが有益であるとの示唆を得ることができた．

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，人々が他と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（あるいは二項，三項間の）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．

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(SEP):1981–2014, 2008.

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

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

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

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

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

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

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

Igarashi, M. and Terui, N. Characterization of Topic-based Online Communities by Combining Network Data and User Generated Content. *Statistics and Computing*, 2020. (forthcoming).

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

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

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

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

Saldaña, D. F., Yu, Y., and Feng, Y. How Many Communities Are There? *Journal of Computational and Graphical Statistics*, 26(1):171–181, 2017.

Snijders, T. A. and Nowicki, K. Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14(1):75–100, 1997.

Wang, Y. J. and Wong, G. Y. Stochastic Blockmodels for Directed Graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

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

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

Tables

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	次数の異質性
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

表 2: WAIC によるモデル比較：K はコミュニティ数を，L はトピック数を表し，太字は最小の値を意味する．

	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

表 3: AUC による予測性能の比較；各モデルの名前はそれぞれ，MMSB が Airoldi et al. (2008)，Homo が Igarashi and Terui (2020)，Hetoro が本研究のモデルを指し，太字は各コミュニティ数（K），トピック数（L）の組み合わせにおける最大の AUC を表す.

	L					
	5	6	7	8	9	10
K=5						
MMSB	0.897	0.897	0.897	0.897	0.897	0.897
Homo	0.896	0.900	0.890	0.883	0.905	0.890
Hetero	**0.917**	**0.920**	**0.924**	**0.921**	**0.923**	**0.924**
K=6						
MMSB	0.913	0.913	0.913	0.913	0.913	0.913
Homo	0.904	0.908	0.910	0.909	0.908	0.906
Hetero	**0.925**	**0.930**	**0.922**	**0.930**	**0.923**	**0.921**
K=7						
MMSB	0.907	0.907	0.907	0.907	0.907	0.907
Homo	0.920	0.900	0.895	0.900	0.903	0.911
Hetero	**0.929**	**0.926**	**0.929**	**0.930**	**0.929**	**0.931**
K=8						
MMSB	0.904	0.904	0.904	0.904	0.904	0.904
Homo	0.925	0.918	0.916	0.921	0.897	0.916
Hetero	**0.928**	**0.930**	**0.927**	**0.930**	**0.929**	**0.928**
K=9						
MMSB	0.912	0.912	0.912	0.912	0.912	0.912
Homo	0.920	0.922	0.918	0.917	0.922	0.920
Hetero	**0.922**	**0.923**	**0.925**	**0.927**	**0.927**	**0.922**
K=10						
MMSB	0.906	0.906	0.906	0.906	0.906	0.906
Homo	0.923	**0.926**	0.923	**0.925**	0.920	0.917
Hetero	**0.923**	0.921	**0.926**	0.925	**0.923**	**0.930**

Figures

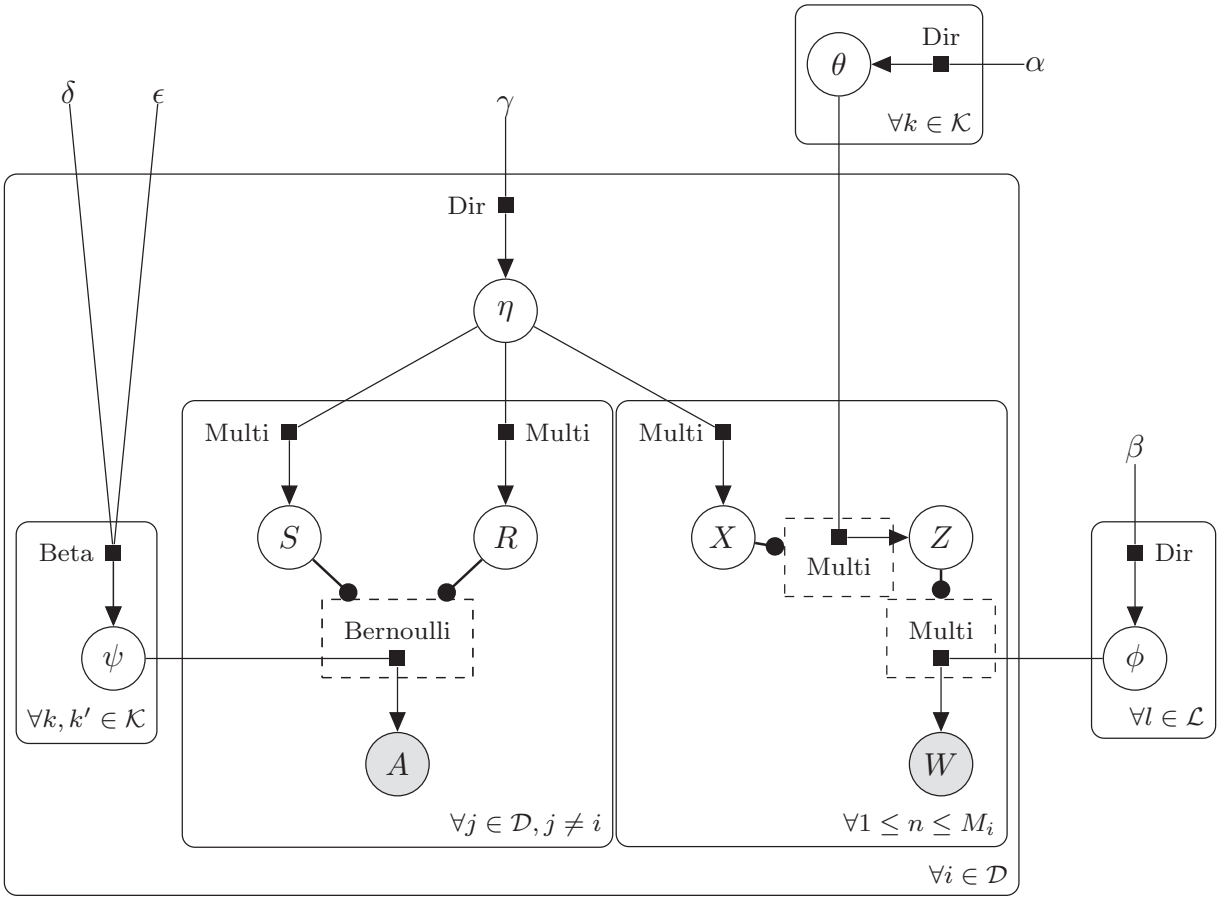

図 1: 提案モデルのグラフィカル表現

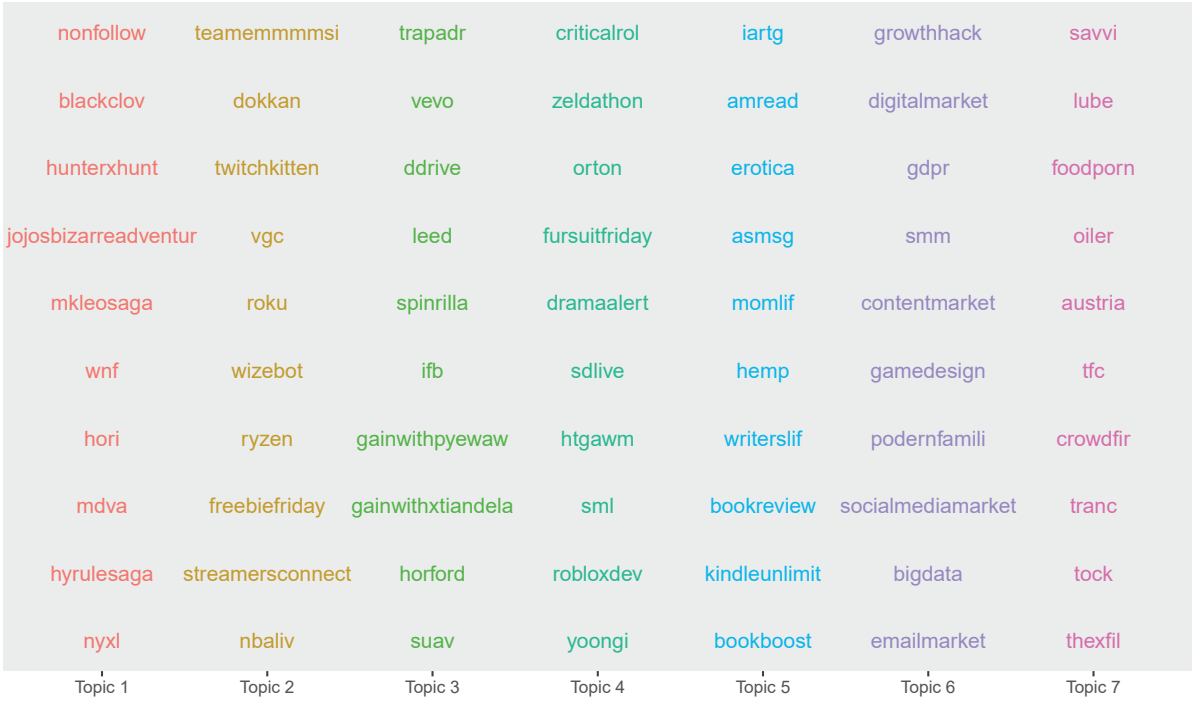

図 2: 推定された単語分布において最も高い値を持つ上位 10 個の単語

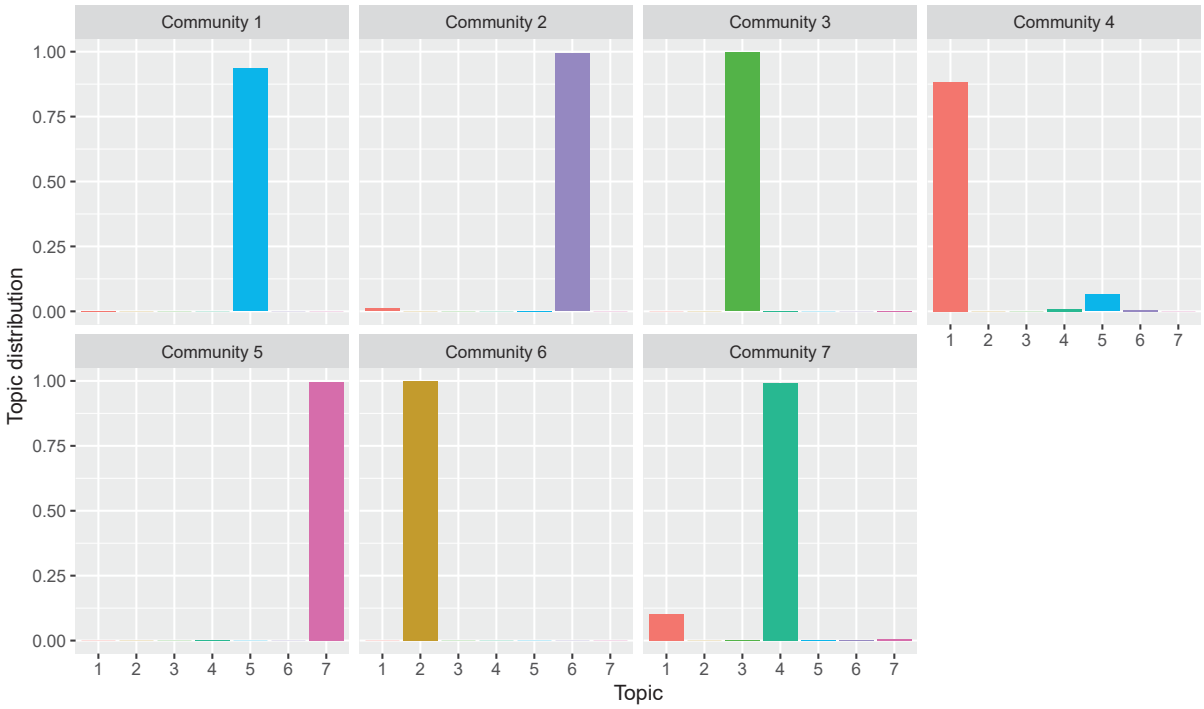

図 3: 各コミュニティのトピック分布に関する推定結果

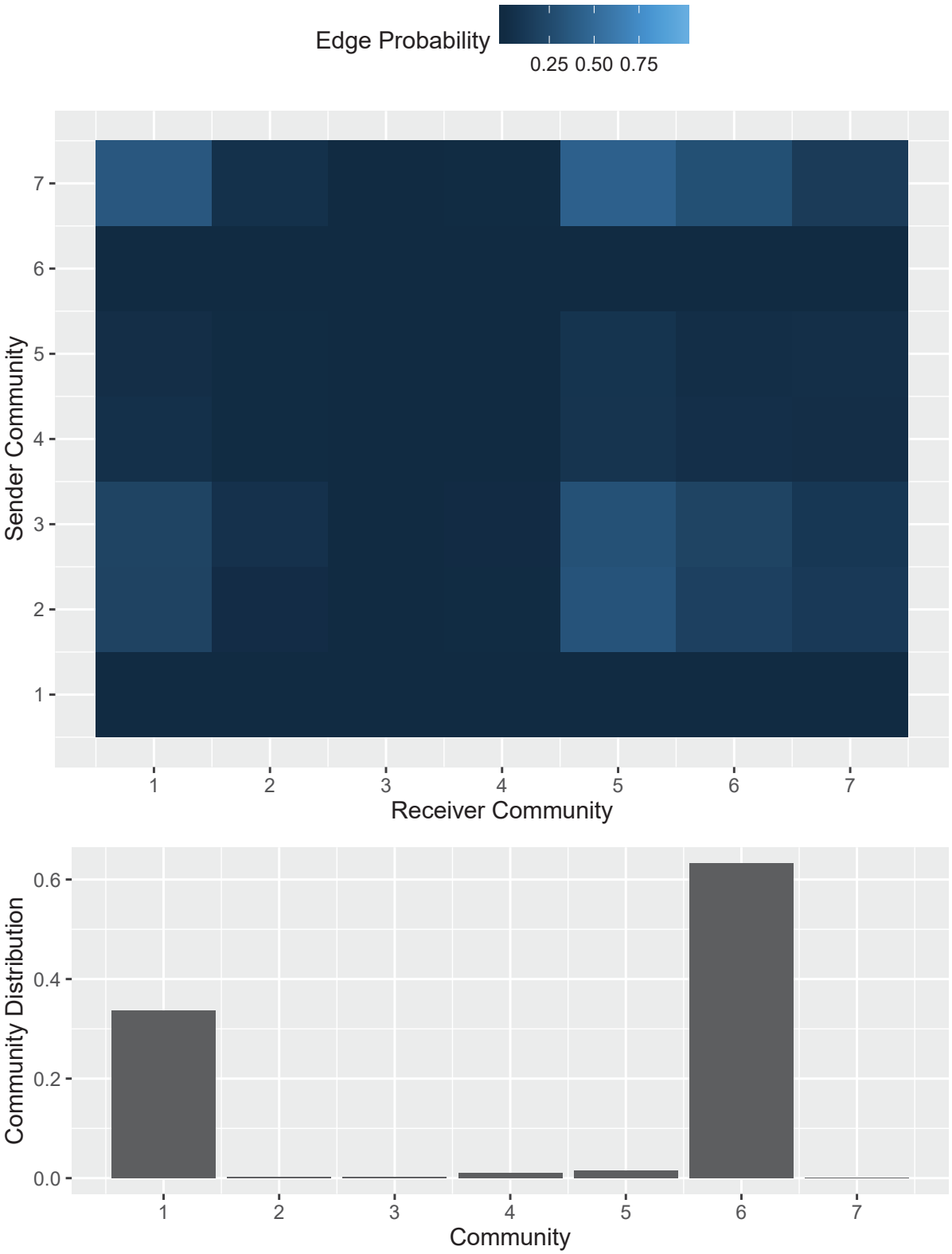

図 4: ノード 1 のエッジ確率とコミュニティ分布に関する推定結果

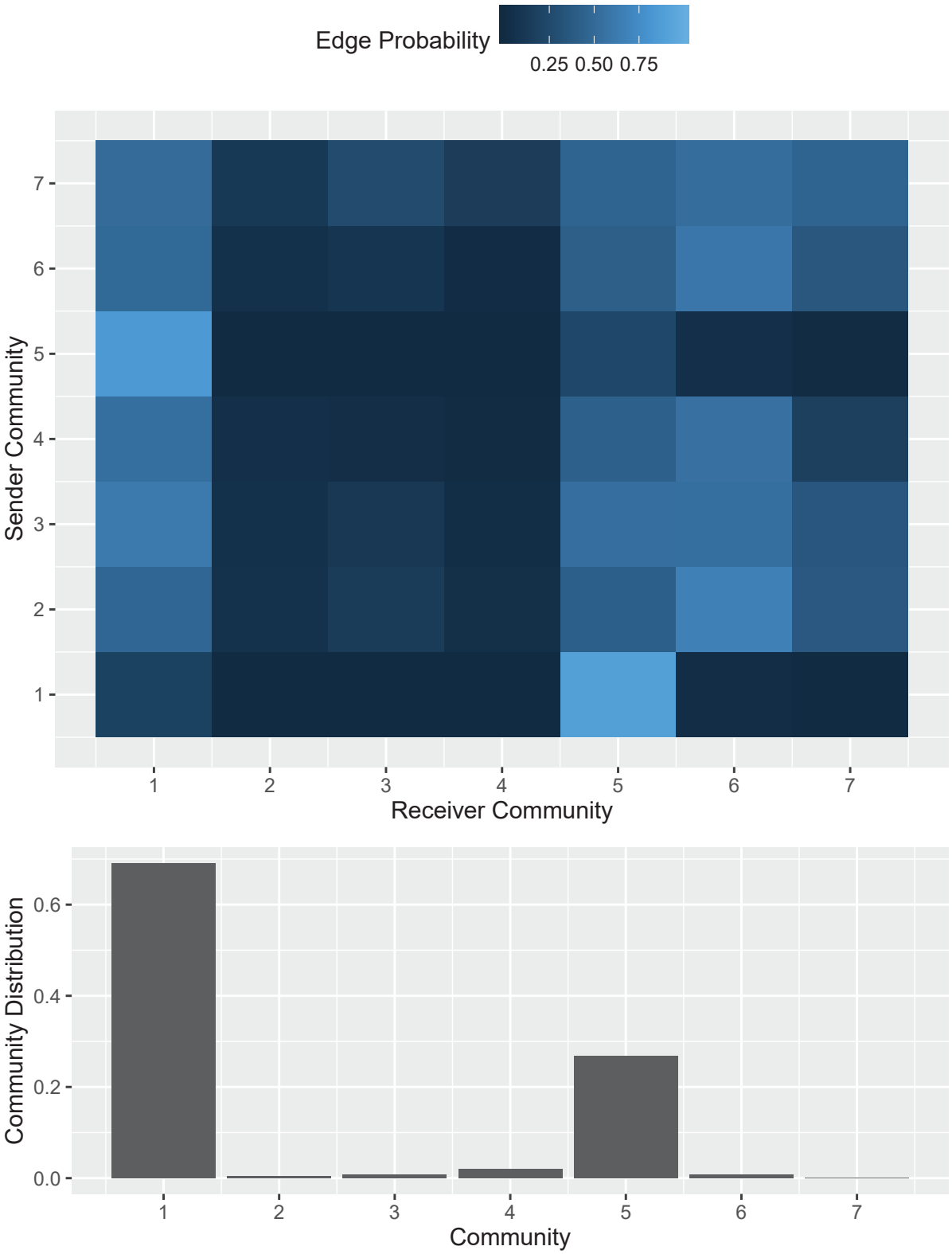

図 5: ノード 237 のエッジ確率とコミュニティ分布に関する推定結果

Appendices

A　条件付き事後分布の導出

??節では，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（式 5，6）．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．

$$p(\eta_i \mid S, R, X, \gamma) = \frac{\Gamma\left(\sum_k N_{ik} + M_{ik} + \gamma_k\right)}{\prod_k \Gamma(N_{ik} + M_{ik} + \gamma_k)} \prod_{k=1}^{K} \eta_{ik}^{N_{ik}+M_{ik}+\gamma_k} \tag{8}$$

$$p(\psi_{ikk'} \mid A, S, R, \delta, \epsilon) = \frac{\Gamma(n_{ikk'}^{(+)} + n_{ikk'}^{(-)} + \delta_{kk'} + \epsilon_{kk'})}{\Gamma(n_{ikk'}^{(+)} + \delta_{kk'})\Gamma(n_{ikk'}^{(-)} + \epsilon_{kk'})} \times \psi_{ikk'}^{\mathbb{I}(a_{ij}=1)}(1 - \psi_{ikk'})^{\mathbb{I}(a_{ij}=0)} \tag{9}$$

$$p(\theta_k \mid X, Z, \alpha) = \frac{\Gamma\left(\sum_l M_{kl} + \alpha_l\right)}{\Pi_l \Gamma(M_{kl} + \alpha_l)} \prod_{l=1}^{L} \theta_{kl}^{M_{kl}+\alpha_l} \tag{10}$$

$$p(\phi_l \mid W, Z, \beta) = \frac{\Gamma\left(\sum_v M_{lv} + \beta_v\right)}{\Pi_v \Gamma(M_{lv} + \beta_v)} \prod_{v=1}^{V} \phi_{lv}^{M_{lv}+\beta_v}, \tag{11}$$

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．

$$lpd^{(i)} = \log\left(\frac{1}{G}\sum_{g=b+1}^{G}\prod_{j=1}^{D} p\left(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}\right) \prod_{m=1}^{M_i} p\left(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right)\right) \tag{12}$$

$$\begin{aligned} p_{waic}^{(i)} = \frac{G}{G-1}\Bigg(&\frac{1}{G}\sum_{g=b+1}^{G}\left(\sum_{j=1}^{D}\log p\left(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}\right)^2 + \sum_{m=1}^{M_i}\log p\left(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right)^2\right) \\ &- \left(\frac{1}{G}\sum_{g=b+1}^{G}\left(\sum_{j=1}^{D}\log p\left(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}\right) + \sum_{m=1}^{M_i}\log p\left(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right)\right)\right)^2\Bigg) \end{aligned} \tag{13}$$

$$WAIC = -2\sum_{i=1}^{D}\left(lpd^{(i)} - p_{waic}^{(i)}\right), \tag{14}$$

ただし，$p\left(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}\right)$ と $p\left(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right)$ 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．

$$p\left(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}\right) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \eta_{ik} \cdot \eta_{jk'}^{(g)} \cdot \psi_{jkk'}^{(g)\mathbb{I}(a_{ij}=1)} \cdot \left(1 - \psi_{jkk'}\right)^{(g)\mathbb{I}(a_{ij}=0)} \tag{15}$$

$$p\left(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right) = \sum_{k=1}^{K} \sum_{l=1}^{L} \eta_{ik}^{(g)} \cdot \theta_{kl}^{(g)} \cdot \phi_{lw_{im}}^{(g)}. \tag{16}$$