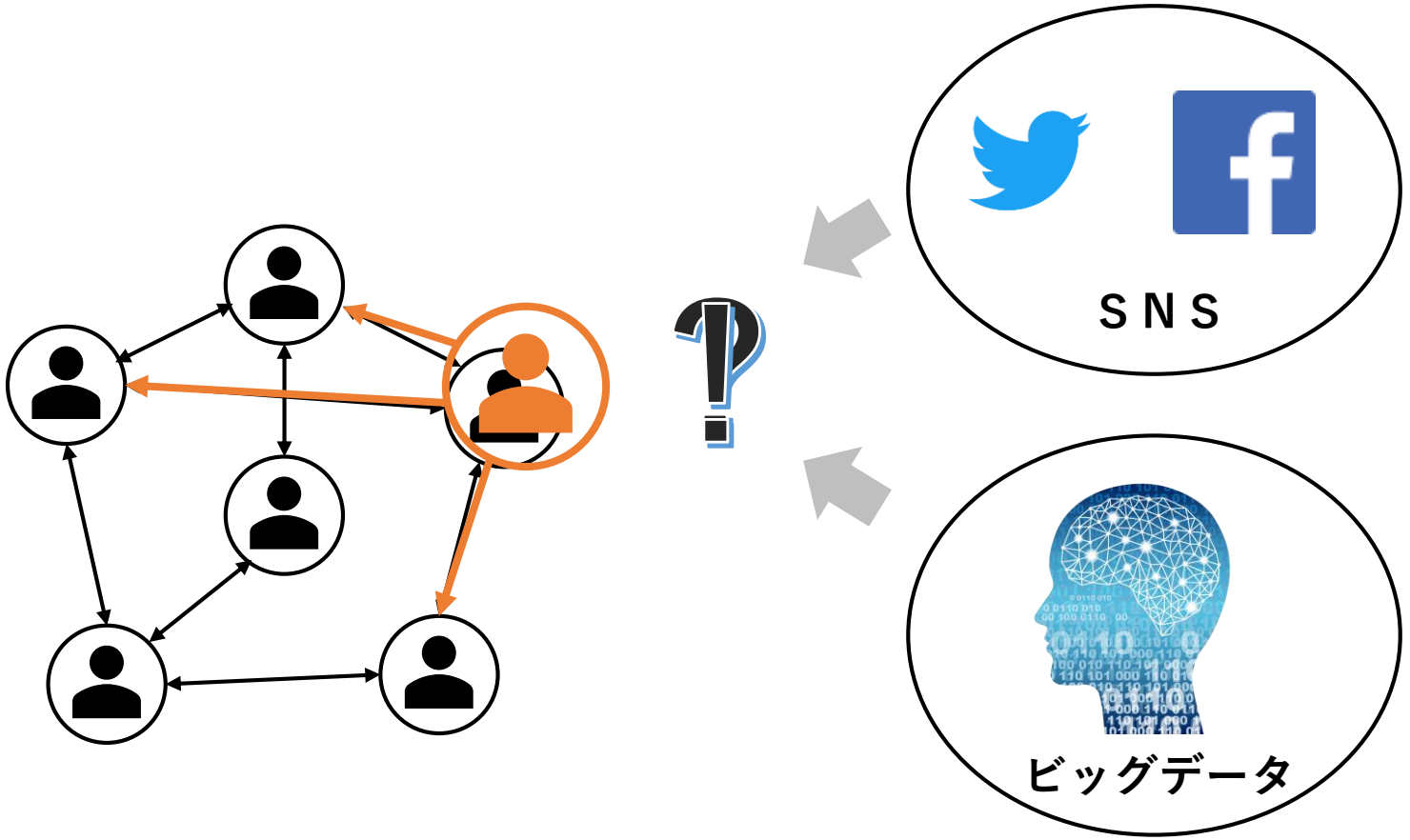
トピックモデルによる ソーシャルネットワーク分析

東北大学大学院　経済学研究科
五十嵐未来
共同研究者：照井伸彦（東北大学）

研究背景



先行研究 (1/3)

- インフルエンサー

  商品の普及や情報の拡散には、ネットワーク上の他者の行動が影響している (Van den bulte & Wuyts 2007)

  その中でもインフルエンサーが果たす役割は大きい

  (Hinz et al. 2011, Iyenger et al. 2011など)

  ネットワーク関係をデータとして与えて中心性により検出する

- ネットワークの扱い

  単層二値ネットワーク (Hinz et al. 2011, Park et al. 2018など) から

  多層ネットワーク (Aral & Walker 2014, Hu & Van den Bulte 2014など) や

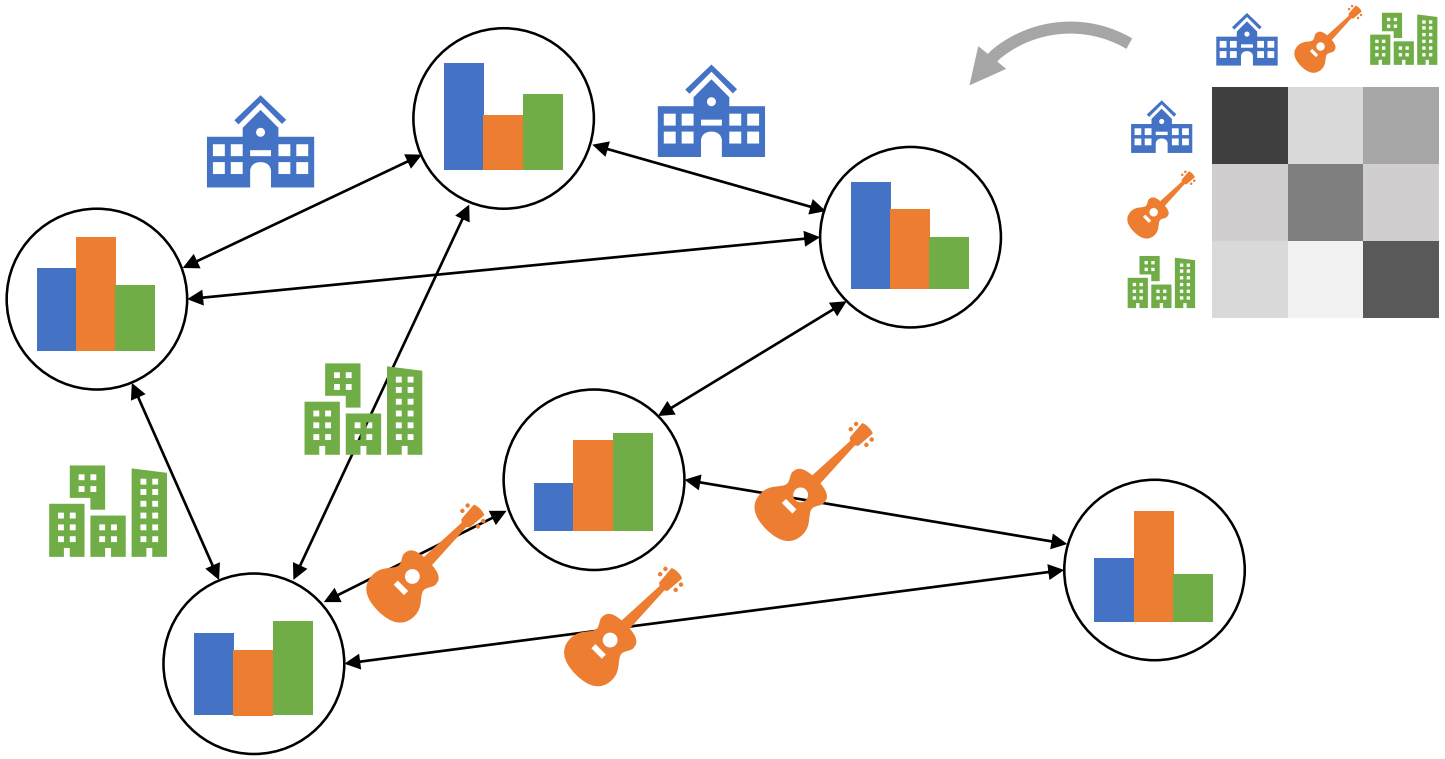
  重み付きネットワーク (Trusov et al. 2010, Chen et al. 2017など) が主流に

  ⇩

  しかし、人々の関係性を調査する必要があり、大規模化が難しい

先行研究 (2/3)

- トピックモデルによるネットワーク分析



先行研究 (3/3)

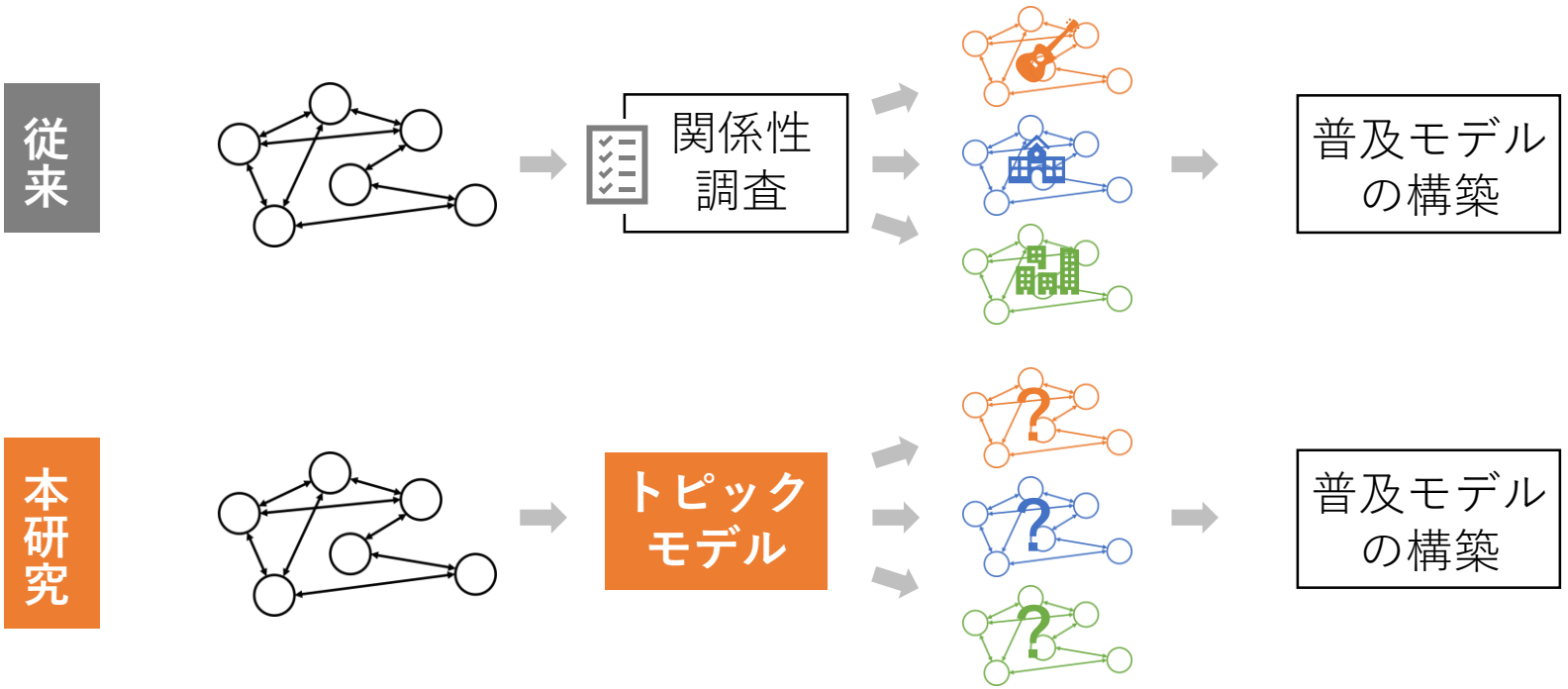
- ネットワーク分析のためのトピックモデル
  - Mixed Membership Stochastic Blockmodels (Airoldi et al. 2008)

    代表的なネットワーク分析のためのトピックモデル
    コミュニティ検出の手法として多くの研究で用いられている

    (Goplan & Blei 2013など)
  - LDA for Group (Henderson & Eliassi-Rad 2009)
  - Simple Social Network LDA (Zhang et al. 2007)　など

本研究ではMMSBを用いる
（→ 他の手法との比較は今後の課題）

研究目的

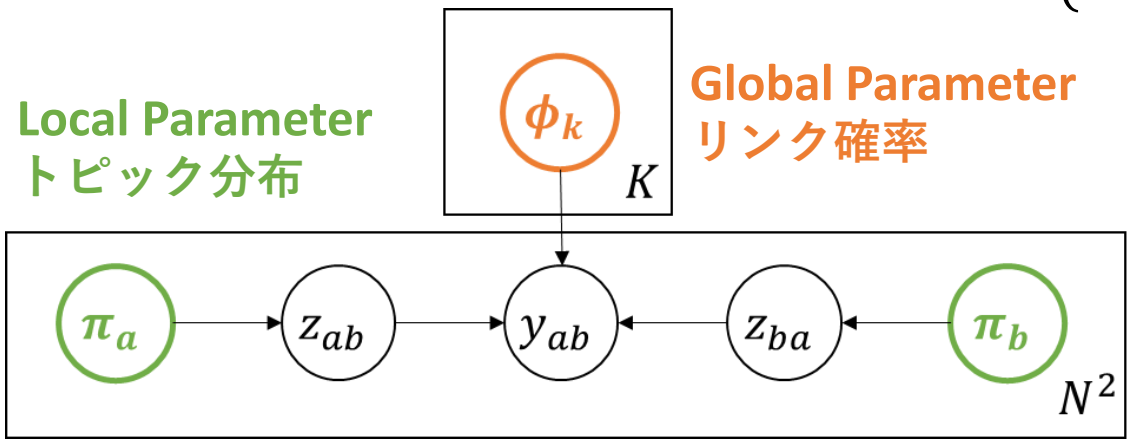


- 研究目的
  - トピックモデルによって人々の関係性を推定することにより、大規模ネットワークに対しても分析可能な普及モデルを提案する
  - 想定していない関係性を抽出できる可能性があり、新たな理論構築・知見創出の一助とする

分析モデル (1/3)

- Assortative Mixed Membership Stochastic Blockmodels
  1. For each topic $k = 1, \ldots, K$ $\quad \phi_k \sim Beta(\eta)$
  2. For each node $a = 1, \ldots, N$ $\quad \pi_a \sim Dirichlet(\alpha)$
  3. For each pair of nodes a and b
     a. Draw interaction indicator $\quad z_{ab} \sim Categorical(\pi_a)$
     b. Draw interaction indicator $\quad z_{ba} \sim Categorical(\pi_b)$
     c. Draw link $\quad y_{ab} \sim Binomial(\gamma)$

$$\text{where, } \gamma = \begin{cases} \phi_k & \text{if } z_{ab} = z_{ba} = k \\ \delta & \text{otherwise} \end{cases}$$



分析モデル (2/3)

- リンク確率

  a-MMSBを推定して得られたリンク確率とトピック分布を用いて、トピック毎のリンク確率$W^k = \{w_{ab}^k\}, k = 1, \ldots, K$を計算する

$$w_{ab}^k = \pi_{ak} \times \pi_{bk} \times \phi_k$$

- 重み付きネットワークの構築

  K種類のネットワークの加重和で重み付きネットワーク $\widetilde{W} = \{\widetilde{w}_{ab}\}$ を構築する (Chen et al. 2017)

$$\widetilde{w}_{ab} = exp\left(\sum_{k=1}^{K} \beta_k w_{ab}^k\right)$$

  β_kは各ネットワークの普及過程への影響を表すパラメータであり、普及データから推定される

分析モデル (3/3)

- 次数中心性

  ネットワークにおけるノードsの次数中心性x_sは次で定義される

$$x_s = \sum_{b=1}^{N} \widetilde{w}_{sb}$$

- 普及モデル

  ノードsによってどれだけの人々に商品が普及したかを示すリーチy_sは、次数中心性x_sと操作変数zによって、次のようにモデル化される (Chen et al. 2017)

$$y_s = \theta_0 + \theta_1 x_s + \boldsymbol{\theta}_2' \boldsymbol{z}_s + \epsilon_s, \qquad with\ \epsilon_s \sim N(0, \sigma^2)$$

推定法

- a-MMSBの推定

  Stochastic Gradient Riemannian Langevin dynamics
  (SGRLD, Li et al. 2015)

  パラメータの勾配をミニバッチで近似しながら更新していく

  → 高速な計算手法でありネットワークの大規模化が可能

  （Collapsed GibbsサンプリングではN^2の関係性全てについて潜在トピックを推定する必要がある）

  ミニバッチ数：15　繰り返し数：1,000回

- 普及モデルの推定

  ネットワークパラメータβ：M-Hサンプリング

  普及モデルパラメータθ, σ：Gibbsサンプリング

  繰り返し数：50,000回（25,000回をBurn-in期間とする）

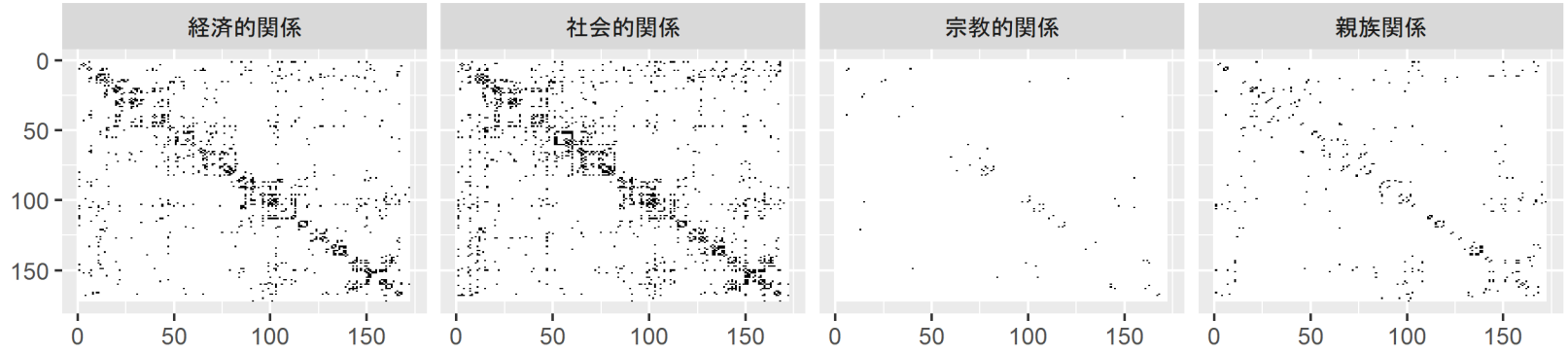
データの概要 (1/3)

- BSSデータ (Banerjee et al. 2013) の概要
  - インドの43の村における小口金融プログラムの普及データ
  - BSSは村の中心人物（教師や商人など）をSeedユーザーと定め口コミの拡散を促した
  - 商品普及に対してネットワークにおける位置が与える影響の分析に適したデータ
- ネットワークデータ

  世帯間の関係性を調査し、4種類の二値ネットワークが得られた
  - 経済的関係（金銭や食料の貸し借りがあったか否か）
  - 社会的関係（医学的な助言や相手宅への訪問などがあったか否か）
  - 宗教的関係（共に寺院へ行ったことがあるか否か）
  - 親族関係（親族関係にあるか否か）

  →対立モデルではすべてのネットワークをデータとして使い、提案モデルでは合算したネットワークのみをデータとして与える

データの概要 (2/3)



		全世帯	Seed 世帯	Non-seed 世帯
世帯数		223.2 (56.2)	26.9 (9.2)	196.3 (50.2)
次数中心性	全体	1.27 (1.00)	1.75 (1.21)	1.21 (.95)
	経済的関係	1.21 (1.00)	1.67 (1.18)	1.15 (.96)
	社会的関係	1.25 (1.00)	1.74 (1.23)	1.17 (.95)
	宗教的関係	.46 (1.00)	.62 (1.19)	.44 (.97)
	親族関係	1.12 (1.00)	1.35 (1.10)	1.09 (.99)

数値は全村の平均値（村数：43）　括弧内は標準偏差

データの概要 (3/3)

- 普及データ

  ある世帯が小口金融施策を受け入れたとき、ネットワーク上で最も近いSeedユーザーが普及に貢献したとみなし、リーチy_sを1増加させる（m人いた場合はm人分のy_sが$1/m$ずつ増加する）

- 操作変数

  Seedユーザーの割合、屋根の形状、部屋の数、電気・トイレ・住宅の形態を操作変数としてモデルに組み込む

	平均	標準偏差	最大値	最小値
リーチ	1.51	1.65	14.86	.00
Seed割合	.00	.03	.07	-.06
屋根（萱）	.29	.45	1.00	.00
屋根（タイル）	.31	.46	1.00	.00
屋根（石）	.21	.41	1.00	.00
屋根（シート）	.16	.36	1.00	.00
部屋数	.00	1.57	15.26	-2.74
電気（形態）	.73	.45	1.00	.00
トイレ（形態）	.39	.50	2.00	.00
家（形態）	.93	.25	1.00	.00

比較モデル

- ネットワークモデル

1. Topic-Networkモデル

   観測した全体ネットワークからトピック毎の関係を推定し
   それらの加重和で重み付きネットワークを構築

$$\tilde{w}_{ab} = exp\left(\sum_k \beta_k w_{ab}^k\right)$$

2. Multi-Networkモデル

   観測した4種類の関係性の加重和で重み付きネットワークを構築

$$\tilde{w}_{ab} = exp\left(\sum_k \beta_k \bar{w}_{ab}^k\right)$$

3. Binary-Networkモデル

   観測した全体ネットワークを二値ネットワークとして扱う

$$\tilde{w}_{ab} = \bar{w}_{ab}$$

- 普及モデル

  全モデルで共通

$$x_s = \sum_{b=1}^{N} \tilde{w}_{sb}\,, \qquad y_s = \theta_0 + \theta_1 x_s + \boldsymbol{\theta}_2' \boldsymbol{z}_s + \epsilon_s, \qquad \epsilon_s \sim N(0, \sigma^2)$$

分析結果

	Binary-Network	Multi-Network	Topic-Network
β_1（経済、Topic 1）	-	-.19 (-.65, .26)	.47 (-2.86, 2.86)
β_2（社会、Topic 2）	-	-.19 (-.69, .36)	-1.66 (-5.70, .99)
β_3（宗教、Topic 3）	-	**.81 (.17, 1.26)**	-2.20 (-6.47, 1.14)
β_4（親族、Topic 4）	-	**1.22 (.83, 1.61)**	**-6.01 (-2.12, -11.29)**
θ_0（切片項）	**.82 (.35, 1.30)**	**.67 (.21, 1.14)**	**2.94 (2.32, 3.55)**
θ_1（次数中心性）	**.08 (.07, .09)**	**.07 (.04, .10)**	**-.007 (-.008, -.005)**
θ_2（Seed割合）	**-11.98 (-14.71, -9.61)**	**-12.30 (-14.96, -9.64)**	**-13.29 (-16.38, -10.28)**
θ_3（屋根・萱）	.05 (-.31, .42)	.12 (-.25, .50)	-.02 (-.44, .41)
θ_4（屋根・タイル）	-.03 (-.40, .34)	-.04 (-.39, .33)	-.09 (-.49, .32)
θ_5（屋根・石）	-.20 (-.59, .17)	-.14 (-.50, .23)	-.39 (-.81, .03)
θ_6（屋根・シート）	-.14 (-.57, .24)	-.17 (-.56, .23)	-.14 (-.60, .32)
θ_7（部屋数）	-.04 (-.11, .01)	-.03 (-.09, .02)	**.10 (.03, .16)**
θ_8（電気・形態）	**-.46 (-.66, -.27)**	**-.41 (-.60, -.22)**	**-.24 (-.46, -.03)**
θ_9（トイレ・形態）	.07 (-.10, .26)	.10 (-.07, .27)	.14 (-.05, .34)
θ_{10}（家・形態）	.09 (-.21, .43)	.09 (-.23, .43)	.17 (-.20, .55)
σ^2（誤差分散）	**1.91 (1.75, 2.08)**	**1.85 (1.70, 2.01)**	**2.39 (2.21, 2.61)**
WAIC	1.75	**1.74**	1.87

数値は事後中央値
括弧内は95%HPD

モデル比較

		Binary-Network	Multi-Network	Topic-Network
In-Sample	R-square	.380	.400	.248
	RMSE	1.270	1.251	1.395
Out-Sample	R-Square	.180	.107	-.702
	RMSE	1.484	1.533	2.055
WAIC		1.75	1.74	1.87

対数周辺尤度 (Chib & Jeliazkov 2001)やWBICによる比較を追加したい（勉強中）

まとめ

- 観測ネットワークからトピック毎のネットワークを推定し、普及過程への影響を測るモデルを提案

- 実証分析ではBinary・Multi・Topic-Networkの三種類で比較
  - 操作変数については、全モデルでほぼ同じ変数が有意
    Topic-Networkモデルでは次数中心性の係数が負に推定
  - ネットワーク変数については、Multi-Networkモデルでは宗教的・親族関係が正に有意
    Topic-NetworkモデルではTopic 4のネットワークが負に有意
  - WAIC比較では、Topic > Binary > Multi-Network

今後の課題

- 提案モデルの精度を改善
  - トピック数の最適化
  - コミュニティ検出ではなくMulti-Network検出
  - 個人パラメータの導入

- オンラインソーシャルネットワークへの応用
  - Twitterにおけるネットワークと商品（情報）の普及（拡散）
  - テキスト情報の利用